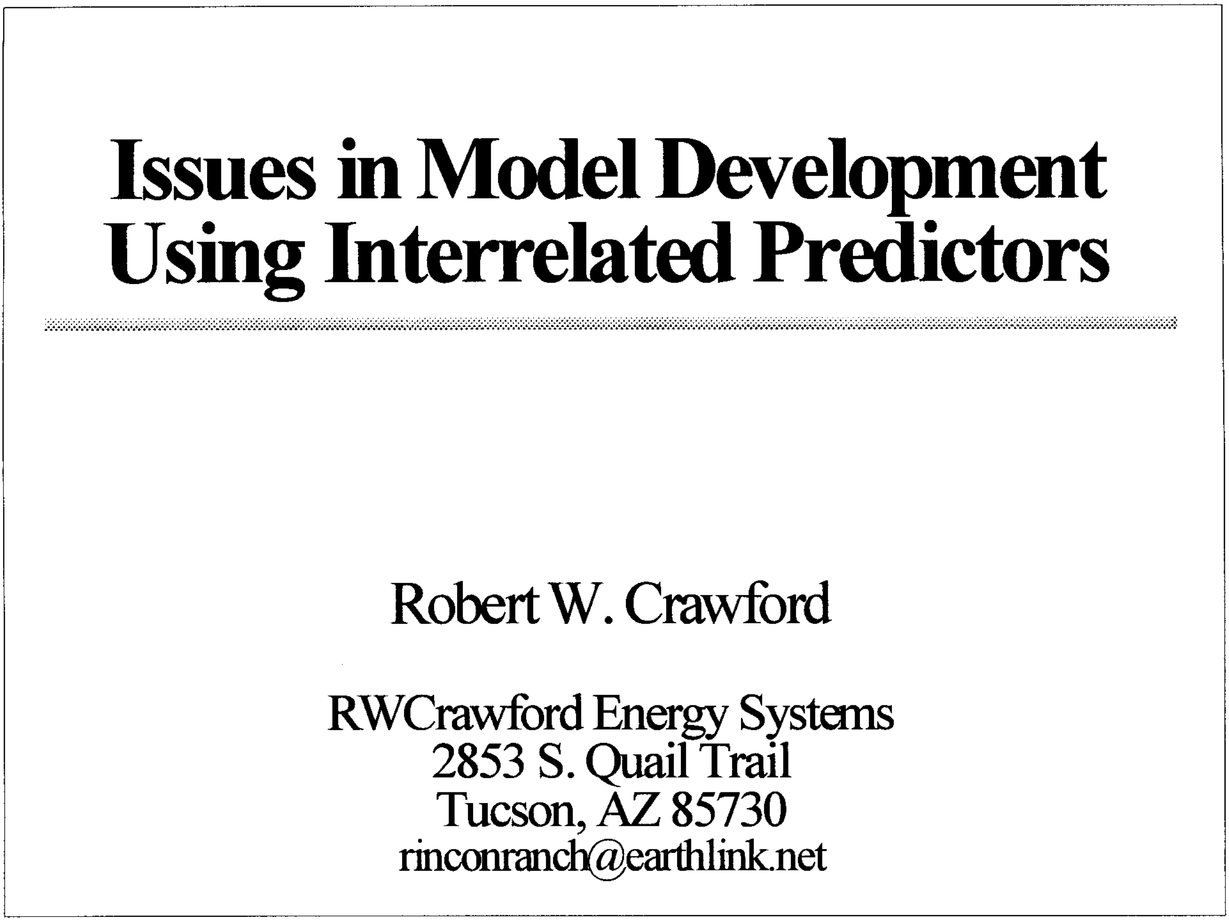
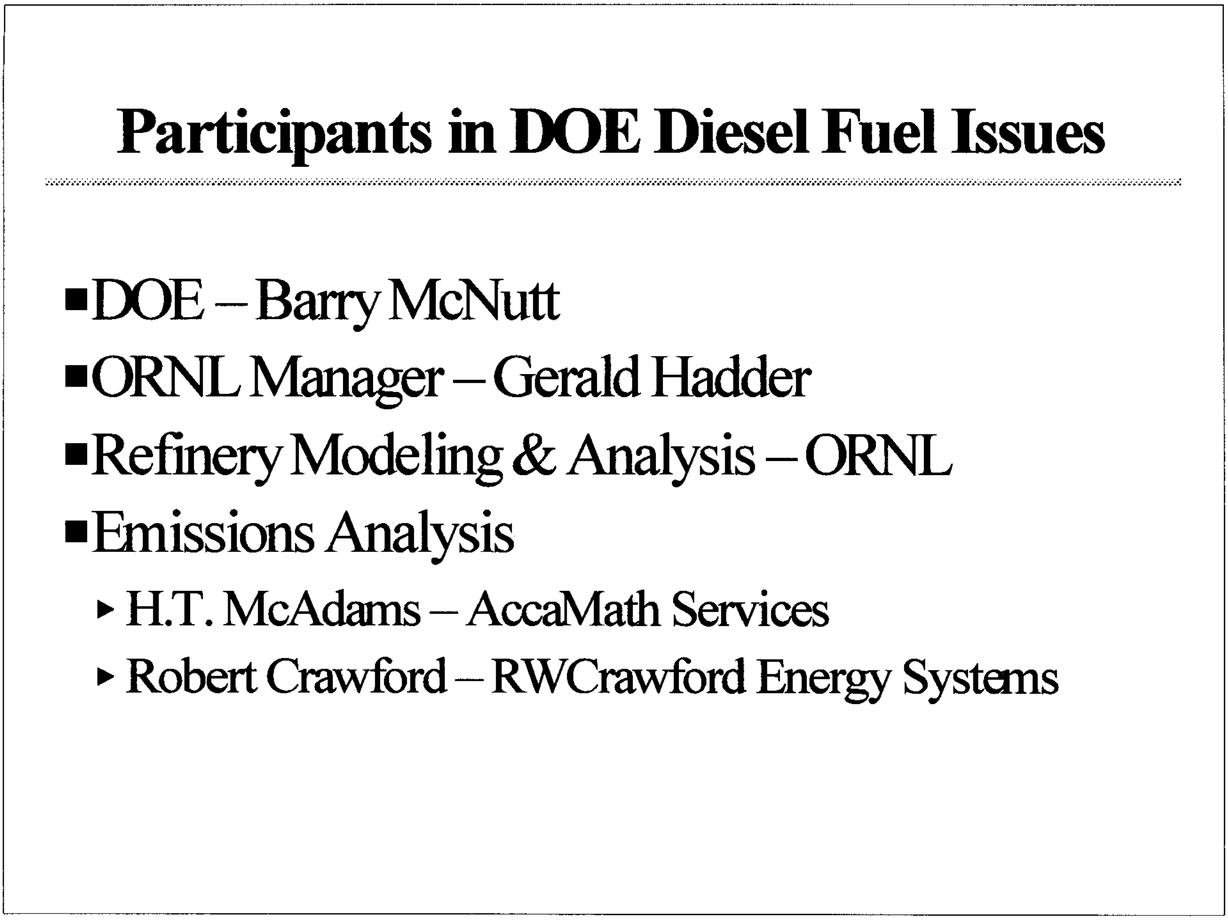


Issues in Model Development Using Interrelated Predictors

Robert W. Crawford

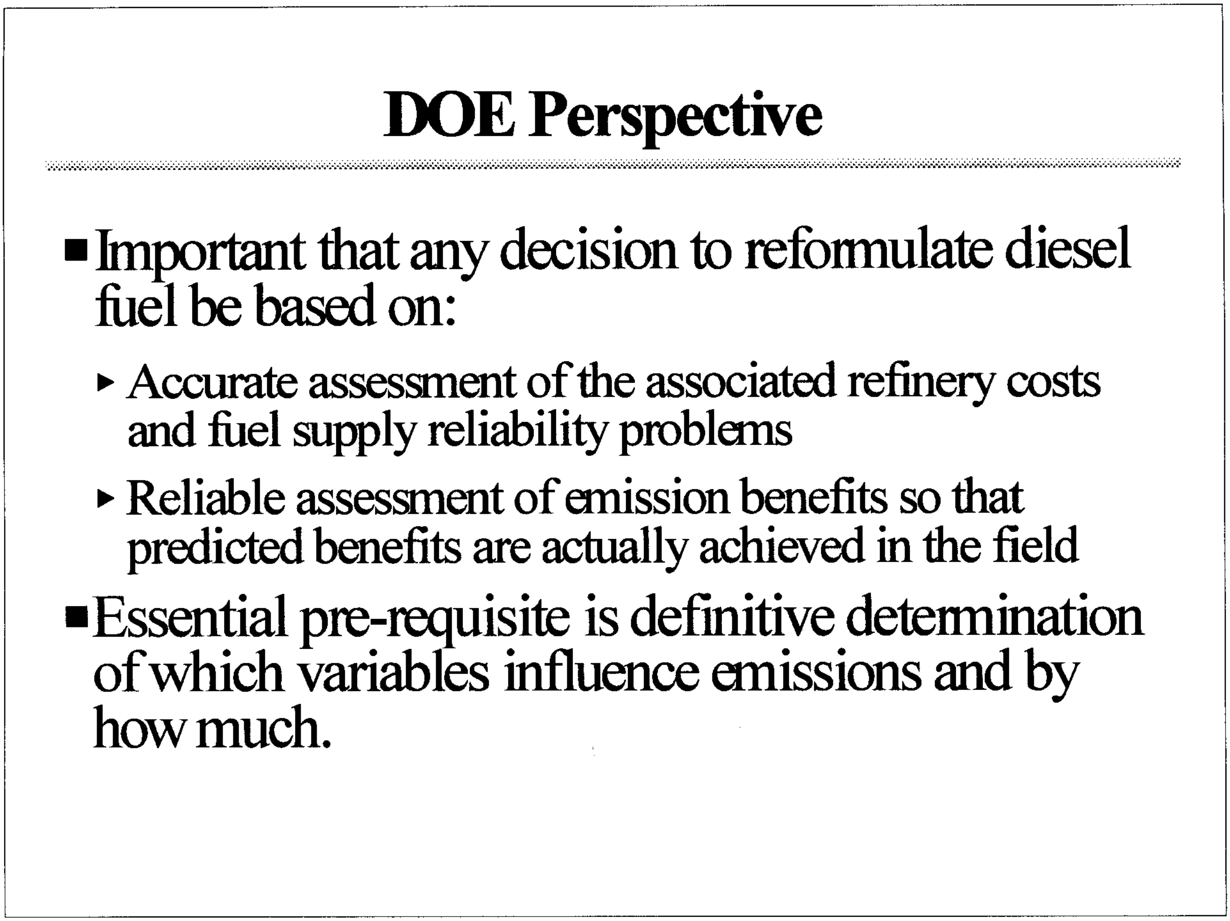
RWCrawford Energy Systems
2853 S. Quail Trail
Tucson, AZ 85730
rinconranch@earthlink.net



Participants in DOE Diesel Fuel Issues

- DOE – Barry McNutt
- ORNL Manager – Gerald Hadder
- Refinery Modeling & Analysis – ORNL
- Emissions Analysis
  - H.T. McAdams – AccaMath Services
  - Robert Crawford – RWCrawford Energy Systems

These are the people involved in our work for DOE on diesel fuel reformulation under the overall guidance of Barry McNutt. Jerry Hadder is in the audience today. HT McAdams had planned to make a presentation today, but was unable to attend the workshop.



DOE Perspective

- Important that any decision to reformulate diesel fuel be based on:
  - ▸ Accurate assessment of the associated refinery costs and fuel supply reliability problems
  - ▸ Reliable assessment of emission benefits so that predicted benefits are actually achieved in the field
- Essential pre-requisite is definitive determination of which variables influence emissions and by how much.

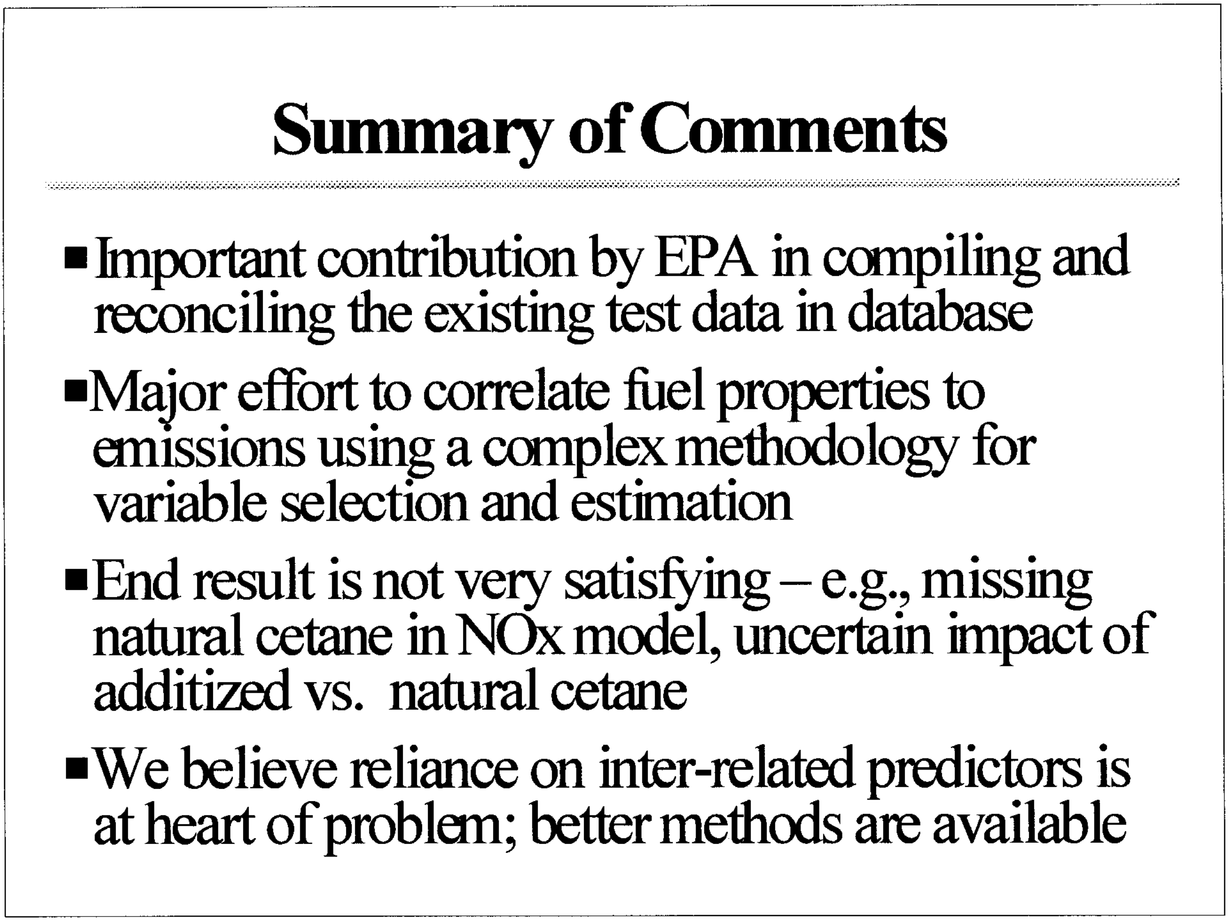
From the DOE perspective, it is important that any decision to reformulate diesel fuel be based on:

o An accurate assessment of the refinery costs associated with reformulation and its impact on fuel supply reliability.

o A reliable assessment of emission benefits that actually will be achieved in the field.

It is the latter of these two points that we will address today.

An essential prerequisite to achieving reliability in the emissions analysis is having a definitive assessment of the variables that influence emissions. If we lack this, there can be no reliable basis for predicting the effects of varying fuel formulation and efforts to implement fuel reformulation may go astray.



Summary of Comments

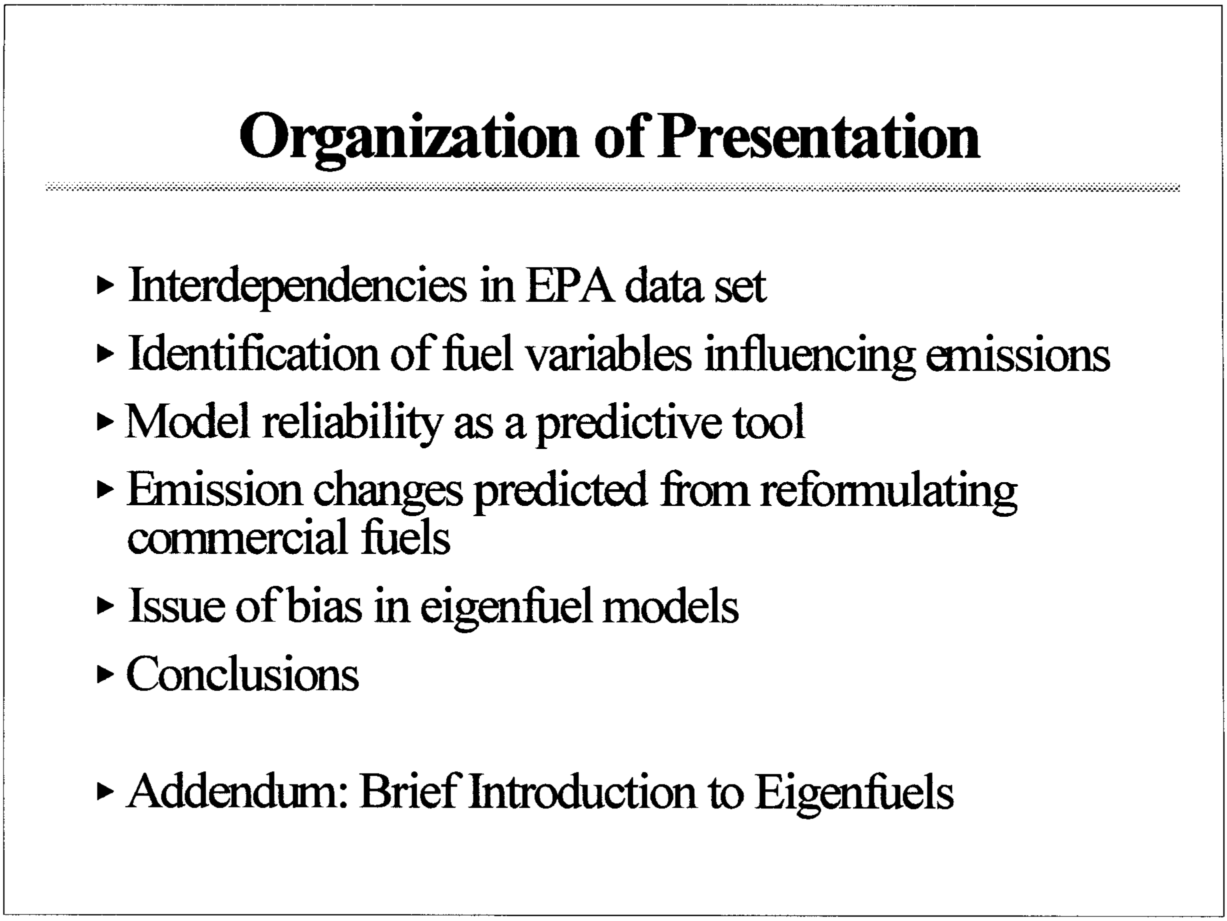
- Important contribution by EPA in compiling and reconciling the existing test data in database
- Major effort to correlate fuel properties to emissions using a complex methodology for variable selection and estimation
- End result is not very satisfying – e.g., missing natural cetane in NOx model, uncertain impact of additized vs. natural cetane
- We believe reliance on inter-related predictors is at heart of problem; better methods are available

Our comments are summarized in this slide. To begin with, EPA has made a very important contribution to this area by the work done in compiling and reconciling the test data that had been previously published. Without this, it has been very difficult to look for a consensus in the test results, and the database will undoubtedly benefit all interested parties.

EPA has made a major effort, in a relatively short timeframe, to understand in detail the correlations between fuel properties and emissions. The methodology for model building is complex, both in the number of individual terms that were considered and how parameters are selected and estimated.

In in the end, however, we find some aspects of the model results to be much less than satisfying -- for example, in the exclusion of natural cetane from the NOx model and the questions regarding the relative impact of natural and additized cetane.

We believe these problems -- in identifying the influential variables and estimating their effects -- stem at least in part from the reliance on inter-related predictors. We have been advocating a different approach called "eigenfuels" as a mean of dealing with these problems directly.

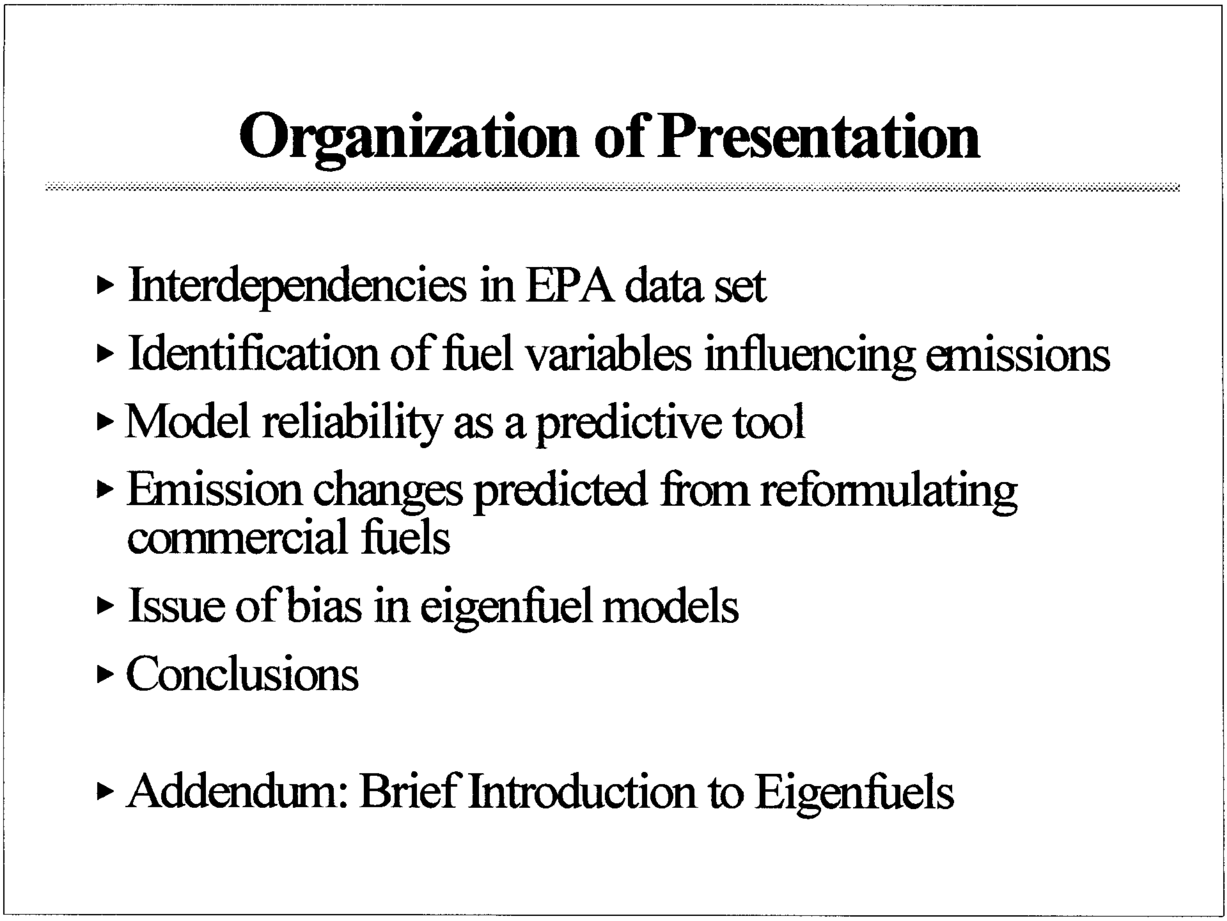


Organization of Presentation

- Interdependencies in EPA data set
- Identification of fuel variables influencing emissions
- Model reliability as a predictive tool
- Emission changes predicted from reformulating commercial fuels
- Issue of bias in eigenfuel models
- Conclusions

- Addendum: Brief Introduction to Eigenfuels

I hope to cover these points in today's presentation. I know that many of you are already familiar with at least some of the eigenfuel work. An addendum on eigenfuels has been included for those of you who have not seen that work. We will be happy to get you more information on eigenfuels, if you would like.



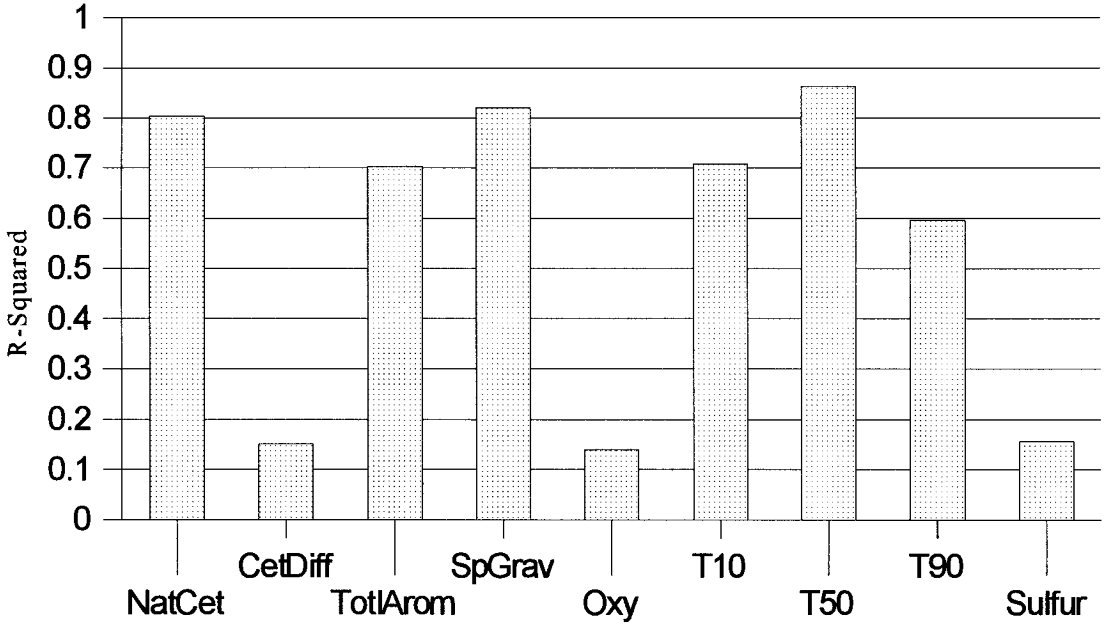
Organization of Presentation

- Interdependencies in EPA data set
- Identification of fuel variables influencing emissions
- Model reliability as a predictive tool
- Emission changes predicted from reformulating commercial fuels
- Issue of bias in eigenfuel models
- Conclusions
- Addendum: Brief Introduction to Eigenfuels

I hope to cover these points in today's presentation. I know that many of you are already familiar with at least some of the eigenfuel work. An addendum on eigenfuels has been included for those of you who have not seen that work. We will be happy to get you more information on eigenfuels, if you would like.

Interdependencies in EPA Data Set

Diesel fuel properties are highly inter-related in the data

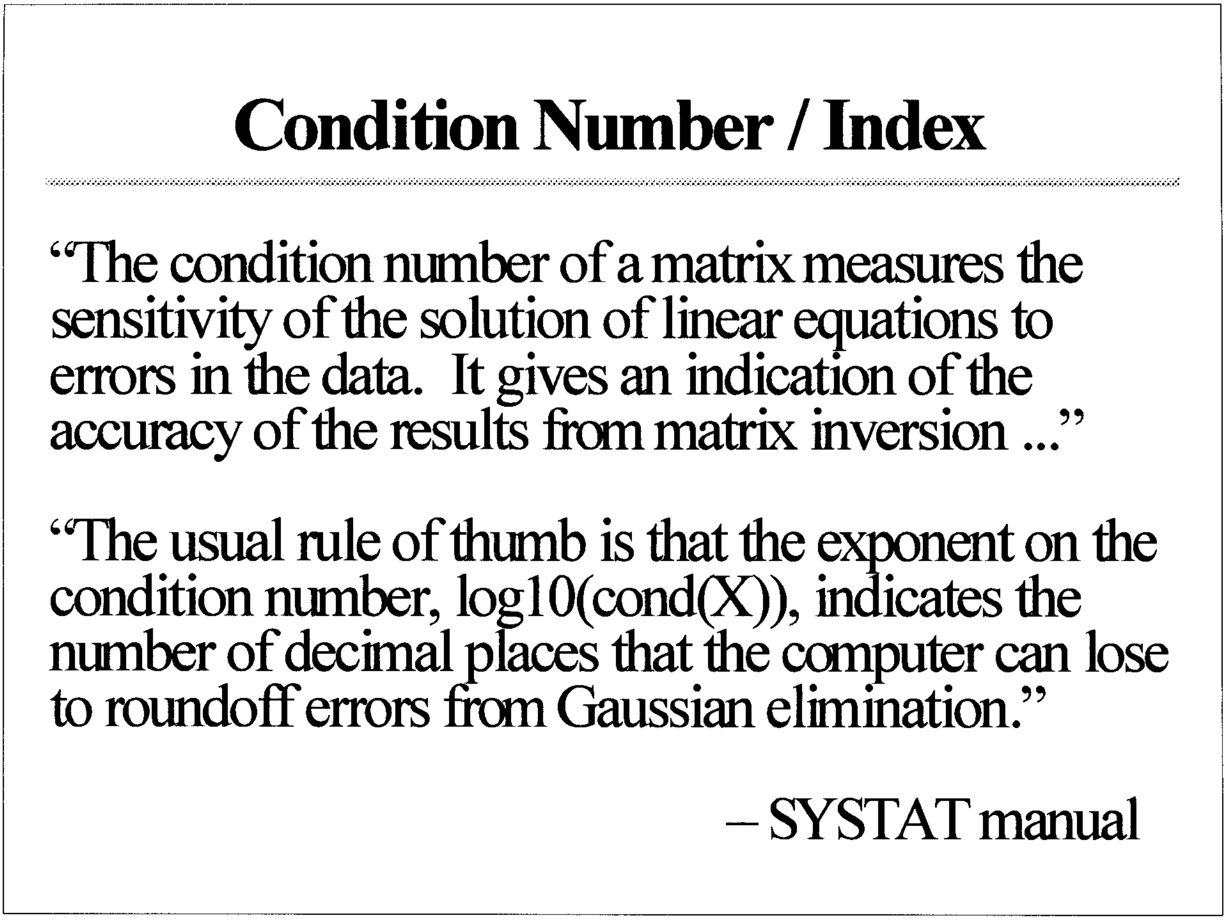


The diesel emissions data set is strongly affected by relationships among the individual fuel properties. As an example, this chart shows the R-squared statistic that is obtained when each property is treated as a response variable and regressed against all other properties.

You can see that only cetane difference, oxygen content and (surprisingly) sulfur content are relatively independent of the other properties. Natural cetane -- distinctive in being a measure of ignition properties and related to physical and chemical factors -- can be largely explained as a function of the other properties. In fact, five of these properties -- natural cetane, total aromatics, specific gravity, T10 and T50 -- can be more than 70 percent explained by the other 8 properties.

This situation gets much worse in the actual model building because both linear and quadratic terms were considered for 3 of the properties. Over the non-negative range of this data, the linear and quadratic terms have correlations on the order of 0.95 with each other. This means that 98 or 99 percent of the terms for natural cetane, cetane difference, and total aromatics can be explained by some other variable.

It should not be a real surprise that the analysis will have trouble distinguishing which of these properties have an effect on emissions and what subset is best used



Condition Number / Index

"The condition number of a matrix measures the sensitivity of the solution of linear equations to errors in the data. It gives an indication of the accuracy of the results from matrix inversion ..."

"The usual rule of thumb is that the exponent on the condition number, log10(cond(X)), indicates the number of decimal places that the computer can lose to roundoff errors from Gaussian elimination."

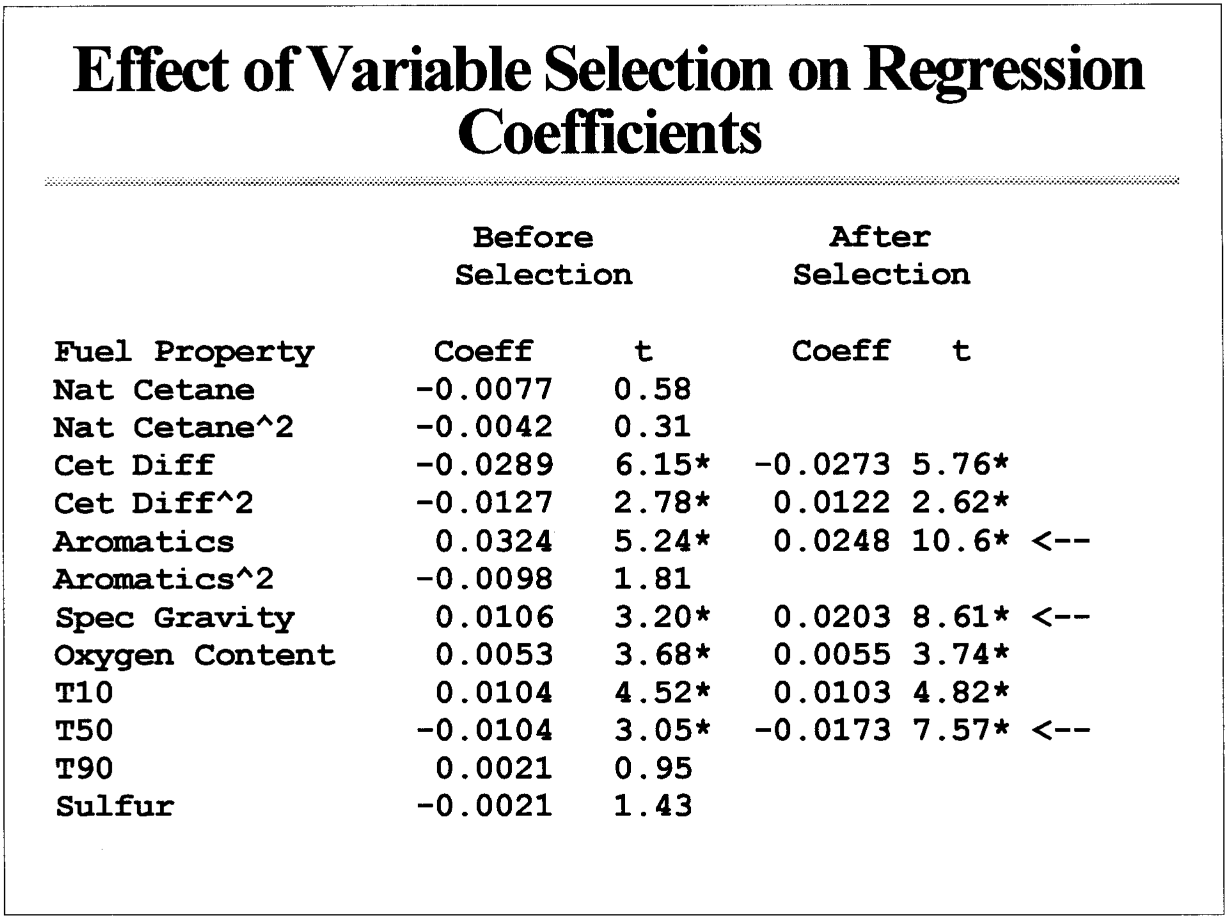
– SYSTAT manual

The EPA analysis concluded its dataset was not materially affected by multicollinearity, because the overall condition number was 5. We view the condition number (or condition index, its square root) as primarily a measure of the computational difficulty faced in the solution of linear equations and the resulting loss of precision. We agree that computational pathology is not a concern in the EPA data set.

However, we do not believe that use of condition number offers an adequate guard against the problems introduced by interdependencies. In fact, any dataset (short of an orthogonal one) has some degree of aliasing among variables and, therefore, some degree of difficulty in identifying unique contributions to the response variable.

Our main target for the eigenfuel approach actually is NOT the set of problems that are computationally pathological -- i.e., situations where the computations will lose precision.

Rather, we are targeting the problem that the predictors are inter-related to such an extent that the analysis will have trouble unraveling their relative importance and individual contributions.



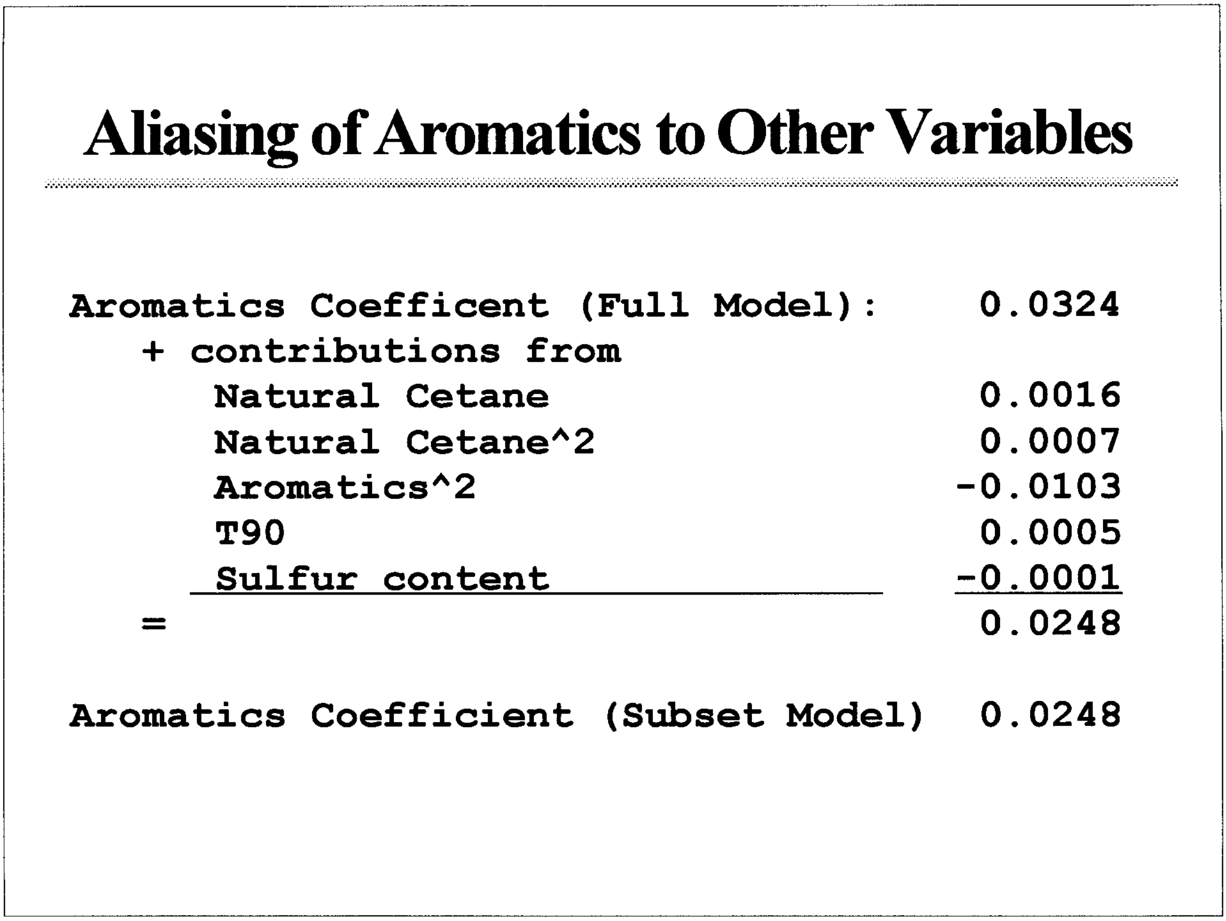
Effect of Variable Selection on Regression Coefficients

	Before Selection		After Selection		
Fuel Property	Coeff	t	Coeff	t	
Nat Cetane	-0.0077	0.58			
Nat Cetane^2	-0.0042	0.31			
Cet Diff	-0.0289	6.15*	-0.0273	5.76*	
Cet Diff^2	-0.0127	2.78*	0.0122	2.62*	
Aromatics	0.0324	5.24*	0.0248	10.6*	<--
Aromatics^2	-0.0098	1.81			
Spec Gravity	0.0106	3.20*	0.0203	8.61*	<--
Oxygen Content	0.0053	3.68*	0.0055	3.74*	
T10	0.0104	4.52*	0.0103	4.82*	
T50	-0.0104	3.05*	-0.0173	7.57*	<--
T90	0.0021	0.95			
Sulfur	-0.0021	1.43			

Let me use an example from the McAdams paper, which will be distributed later, to illustrate the effect of variable selection on the values estimated for the regression coefficients in an environment of inter-related predictors. The data here are for technology group T and have been adjusted to remove the mean emissions level for each engine before regressing on fuel properties.

Here is an ordinary regression in which all 12 of the fuel property variables have been included. As is usually done, the variables satisfying the 0.05 significance level are retained; these are the ones that have been starred. All other terms are rejected. In this case, 7 predictors are retained and 5 are rejected.

I call your attention to three of the retained variables -- aromatics, specific gravity, and T50 -- which are 3 of the 4 variables included in the EPA Unified NOx model. Their coefficients change appreciably when the model is re-estimated with the 7 variables, and they also appear to be more significant in the subset model than in the full model. Understanding why this happens will tell us a lot about the interpretation and reliability of models based on these variables.

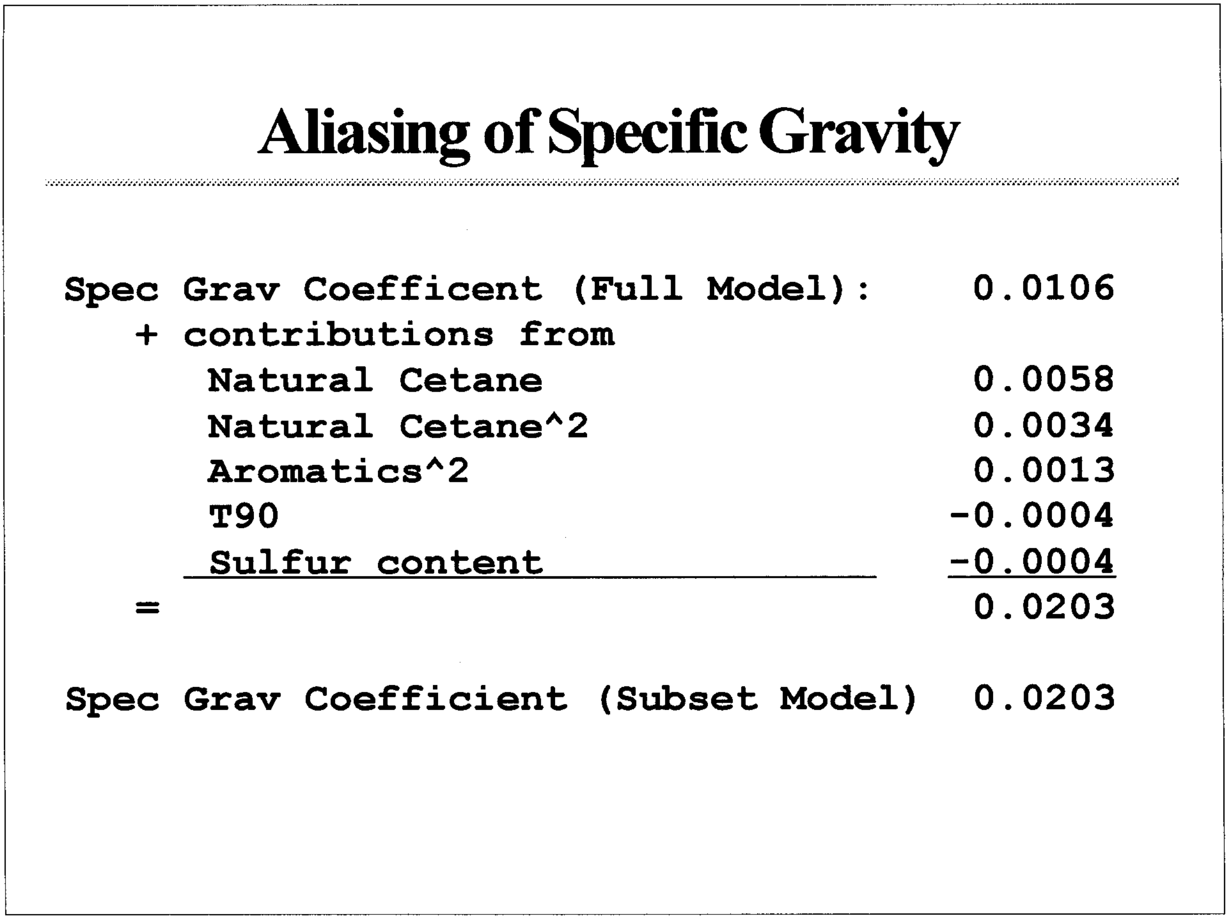


Aliasing of Aromatics to Other Variables

Aromatics Coefficent (Full Model):	0.0324
+ contributions from	
Natural Cetane	0.0016
Natural Cetane^2	0.0007
Aromatics^2	-0.0103
T90	0.0005
Sulfur content	-0.0001
=	0.0248
Aromatics Coefficient (Subset Model)	0.0248

There is a way to decompose the coefficients estimated for a subset model into: (1) the part that originates with each variable retained in the model; and (2) the part(s) that are "picked up" from other, excluded variables with which the retained variables are aliased. The method involves computation of the "alias" or "bias" matrix. This is discussed further in the McAdams paper.

When applied to the coefficient estimated for aromatics content in the subset model, we find that the term includes contributions from five other variables. The largest aliased contribution is from the quadratic term in aromatics content, which has a high correlation to the linear term. But other variables, including natural cetane, T90 and sulfur content also contribute.



Aliasing of Specific Gravity

```
Spec Grav Coefficent (Full Model):    0.0106
   + contributions from
    Natural Cetane                    0.0058
    Natural Cetane^2                  0.0034
    Aromatics^2                       0.0013
    T90                              -0.0004
  _ Sulfur content___________        -0.0004
   =                                  0.0203

Spec Grav Coefficient (Subset Model)  0.0203
```

We see a second example here for specific gravity. Its coefficient nearly doubles between the full and subset regressions. The reason is that it predominantly picks up the predictive contributions of the natural cetane variables (linear and quadratic terms). Other excluded variables make smaller contributions. What we originally thought was the effect of density, is now aliased to include the effects of natural cetane and its square.

This comes about because the predictors are related to each other. The subset model really includes contributions from all the variables, whether those contributions are separately identified or not, and the variables included in the model stand for more than just themselves.

Further, the coefficients in the subset model may be said to be "biased" relative to the full model. This bias will be small only when the terms excluded have negligible effect on the response variable. The bias will be large whenever the model excludes, for any reason, aliased terms that carry an appreciable effect on the response.

Indentifying Influential Variables

Data & Analysis

- 10 technology groups showing common response to fuels for NOx, PM, HC in Unified Model (dominated by T and F groups)
- 906 emission tests out of 1315 in database
- 12 fuel property variables (9 properties + 3 quadratic terms)
- NOx and PM
- OLS estimation, with controls for individual engine effects, but not engines x fuels

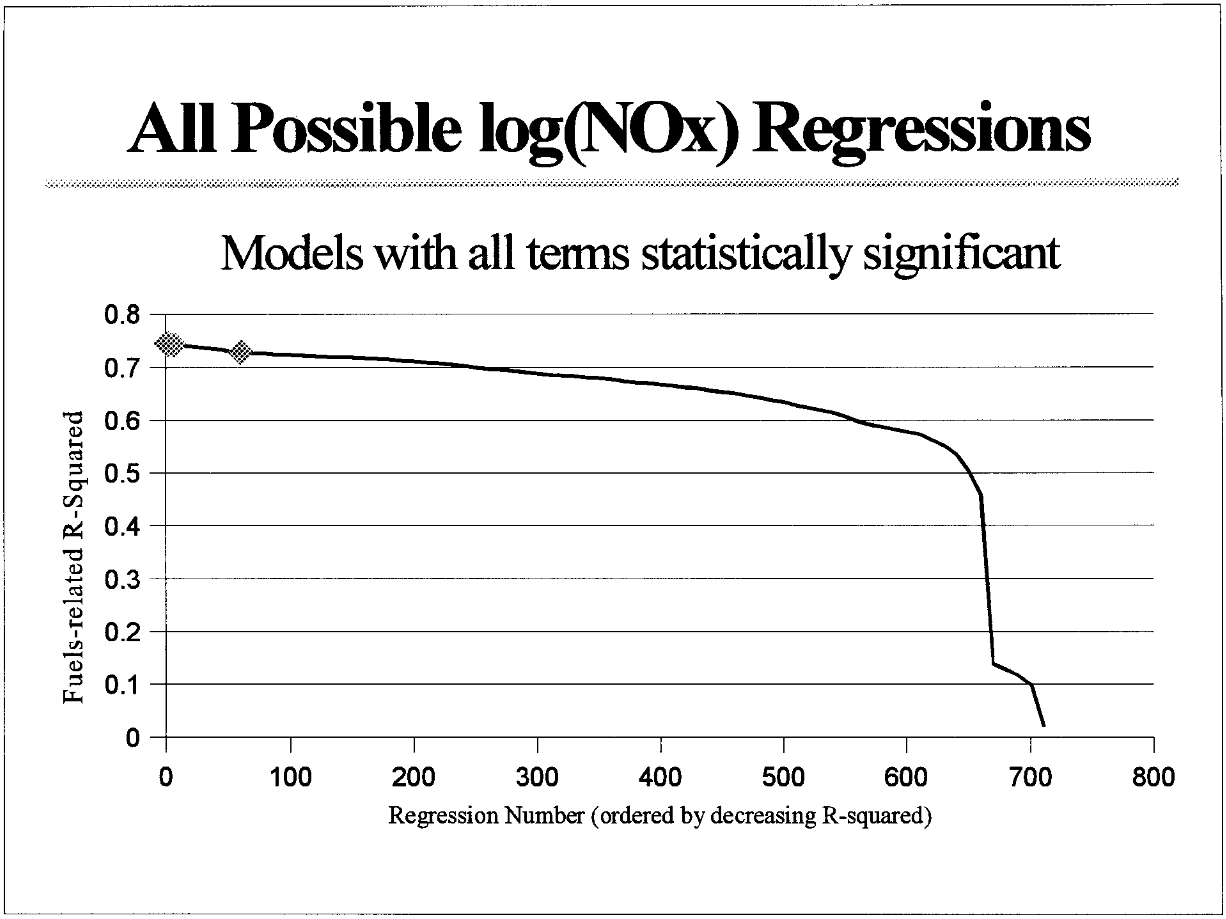
Our comments are largely focused on the effect of the analysis methodology. We have used a consistent subset of the data and selection of variables throughout the analysis we present here, so that comparisons are not influenced by differences in data.

The data subset covers the 10 technology groups that EPA found to share a common response to fuels across pollutants in the Unified Model. This amounts to just under 70 percent of the whole database.

We have also focused on the 12 fuel variables used by EPA -- nine different properties and three quadratic terms -- and just NOx and PM. The presentation shows results for NOx only.

The models are estimated using ordinary least squares (OLS) with controls for the effect of individual engines on emissions. We have not attempted to estimate models with the much larger number of controls for engines x fuels (up to 41 engines x 9 properties in this data subset).

The models presented here do not necessarily represent a judgement of what is the "best model" for DOE purposes. In particular, we need to give consideration to a wider range of non-linear and fuel-fuel interaction terms. We also need to decide how to account for technology group differences.



All Possible log(NOx) Regressions

There are 4095 different regressions that can be formed from 12 fuel variables; each regression also contains controls for individual engine effects. In this instance, at least, it is possible to estimate each of these models and then select the 711 in which all included fuel terms are indicated to be statistically significant at the 0.05 level.

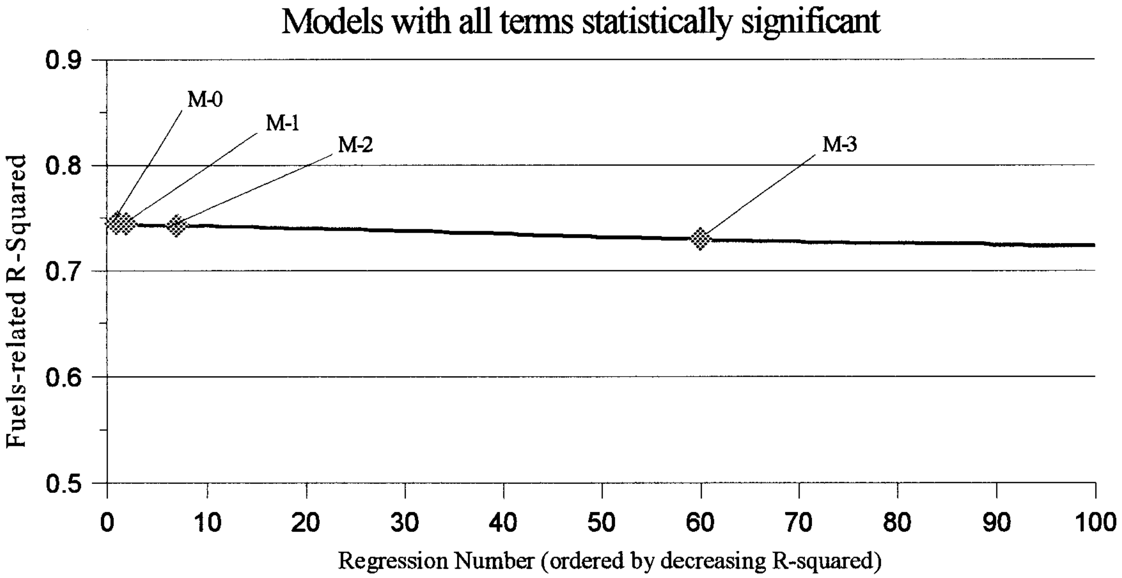
The "fuels-related" R-squared shown in the figure is the R-squared of the fuel variables based on the SS remaining once the explanatory power of the engine control terms has been removed.

The sharp rise to the right at model 664 is the first entry of natural cetane as a variable. By itself as the only fuel variable, natural cetane produces a fuels-related R-squared of 0.42 for log(NOx).

The curve of increasing R-squared bends over at about model 620 and then begins a slow increase to its maximum of 0.745.

All of the interesting action takes place at the left, so let's zoom in closer.

All Possible Regressions: log(NOx)



First, there is not much to choose from among the four models that are highlighted, as long as our focus is on R-squared. The highest R-squared is 0.745 for model M-0 and the lowest is 0.723 at the far right.

M-0. This is the "best model" possible using the 12 variables, in that it achieves the highest R-squared among the models that have statistically significant coefficients. It includes 9 fuel property terms.

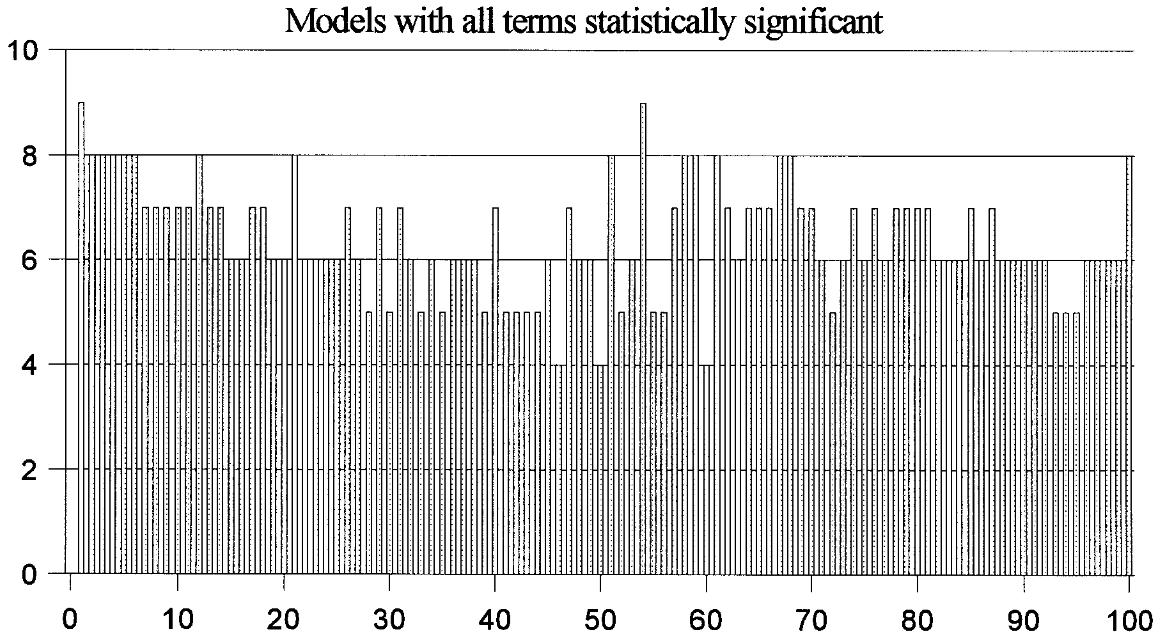
M-1. This is actually the "best model" found using stepwise regression. It is model number 2 and contains 8 fuel terms. This illustrates the caution often given to stepwise regression -- the search process will not necessarily find the "best" model.

M-2. EPA's Unified Model was estimated by a method that includes controls for engine x fuel effects. One aspect of this is that the engine x fuel controls "use up" a number of degrees of freedom. M-2 is the "best model" found by stepwise regression -- among models with engine controls and fuel effect terms, but not interactions -- when the degrees of freedom are reduced by the corresponding amount by deleting data. This is model number 7 and contains 8 terms.

M-3. This is NOT an outcome of stepwise regression, but merely the 4-term model containing EPA's choice of variables in the Unified NOx model (cetane difference, total aromatics, specific gravity, and T50). It has been re-estimated using the dataset with reduced degrees of freedom.

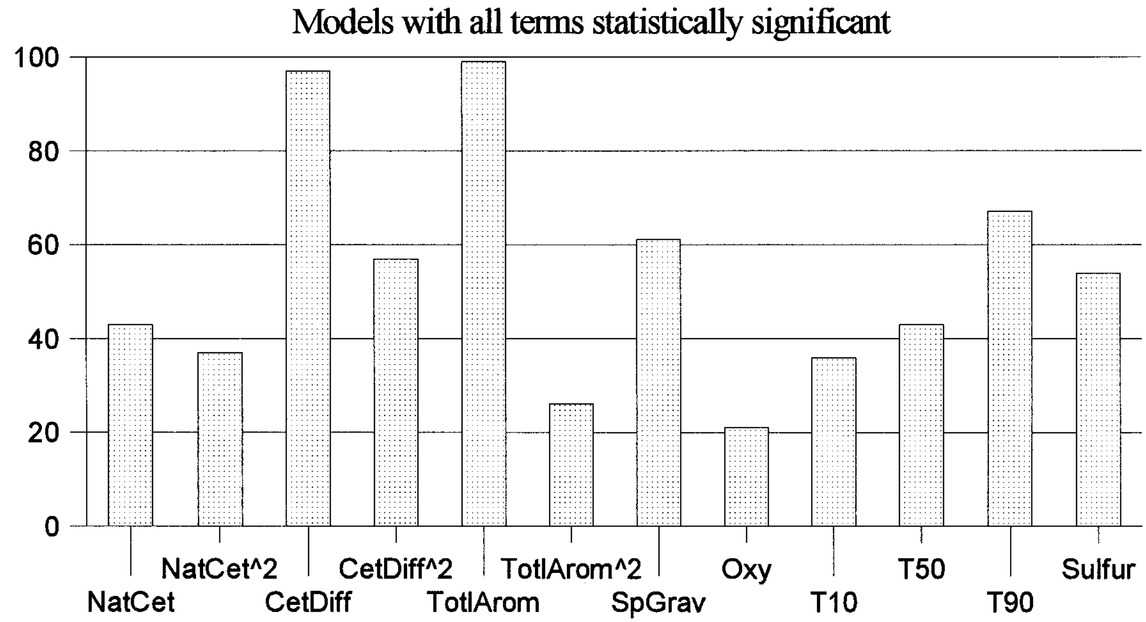
There are other factors, which we cannot simulate here, that contribute in EPA's analysis to the difference between models M-2 and M-3. Those are: the inclusion of engine x fuel interaction terms and the use of a Mixed Effect model solved by an iterative maximum likelihood method

Number of terms in log(NOx) Regressions



This chart shows that the number of fuel terms can vary anywhere from a low of 4 to a maximum of 9 among the 100 best models -- with almost no difference in R-squared.

Frequency of Terms in 100 Best Models



Perhaps more to the point on variable selection, this chart shows the frequency with which the 12 fuel property variables are included in the 100 best models. We certanely believe that cetane difference and total aromatics are important variables, because they are included in all but a couple of models.

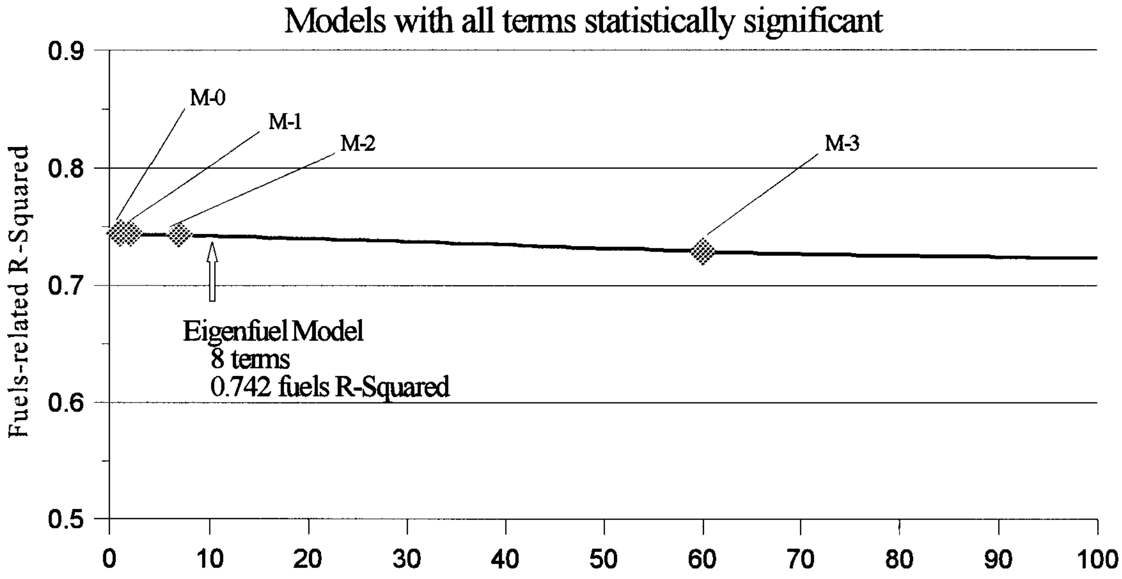
The case is a little less clear for other variables. For example, cetane difference ^2, specific gravity, T90, and sulfur are contained in more than half of the models, while four other variables (natural cetane, natural cetane^2, T10, and T50) are included in a large minority of the models.

The problem we have with stepwise-style variable selection using inter-related predictors is the large number of "good" models you can find that differ in the number and selection of fuel terms. There is little reason to prefer one of these over another based on R-square. The absence of a variable from the model can be due to the fact that its effects are accounted for by other variables present in the model, and not because the variable has no effect. Relatively small differences in the data can lead the analysis to select a different model with a different set of predictors.

The eigenfuel method is a way to deal directly with the problem of aliasing, its effect on variable selection, and the biasing impact on coefficient estimates. We believe that eigenvectors are the only approach capable of providing an

Reliability of Models as Predictive Tools

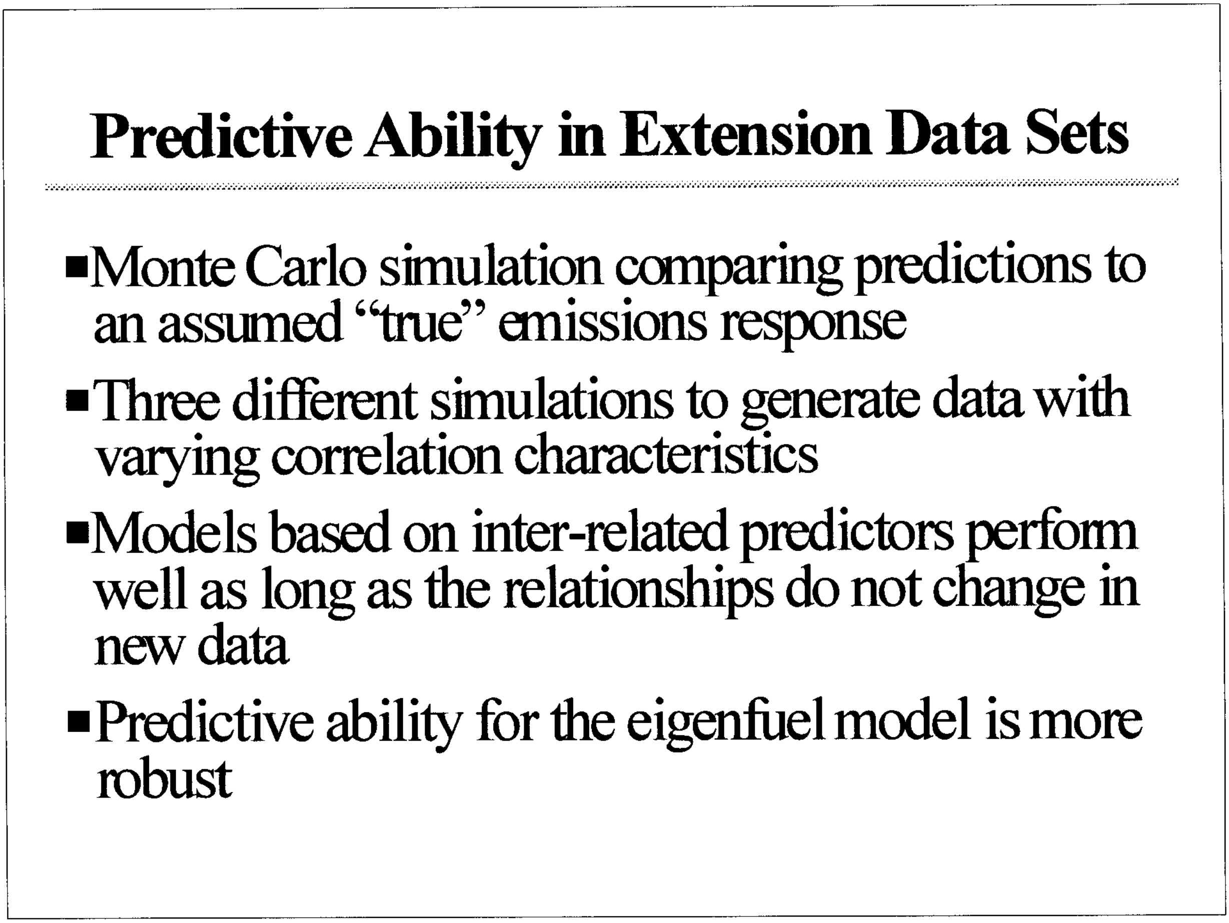
Predictions of Stepwise and Eigenfuel Models



This chart shows us where the eigenfuel model would fall in the range of stepwise models. The eigenfuel model for log(NOx) contains a total of 8 terms that are statistically significant at the 0.05 level. It is estimated with controls for engine effects, but not engine x fuel effects. Its fuels-related R-square places it at position 10 on the R-square curve for the stepwise models. Its 8 eigenvectors contain, of course, all twelve fuel variables.

We can say that there is little basis to prefer any one of these five models -- stepwise or eigenfuel -- based on R-squared. But, R-squared measures only how well the model conforms to the data in the dataset, and it does not assure good prediction except for these points.

We will see next that this situation changes when we examine other datasets. A useful terminology is that the data set used to estimate the model is the "training" data set, and other data sets to which the model is applied are "extension" or "test" data sets. Good predictive ability refers to how successful the model is in making predictions for cases that are not contained in the "training" data set. That is what we mean by "extension".



Predictive Ability in Extension Data Sets

- Monte Carlo simulation comparing predictions to an assumed “true” emissions response
- Three different simulations to generate data with varying correlation characteristics
- Models based on inter-related predictors perform well as long as the relationships do not change in new data
- Predictive ability for the eigenfuel model is more robust

We have looked at the predictive ability of these models when applied to extension data sets using Monte Carlo simulations. Basically, we assumed a "true" emissions response, generated synthetic data with different degrees of correlation among the fuel properties, and then determined how well the several models performed in making predictions.

Three different simulations were done:

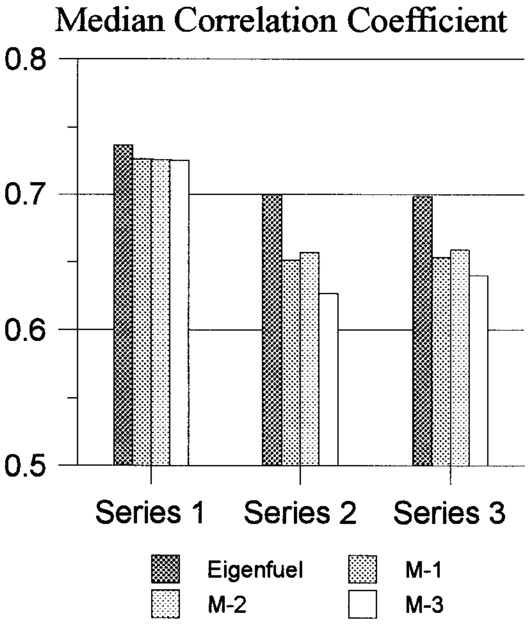
o Series 1: synthetic data having correlations among fuel properties that are very similar to those found in the EPA data set

o Series 2: synthetic data created to have different correlations among properties than those found in the EPA data set

o Series 3: synthetic data in which the fuel properties are completely independent

Predictive Ability of the Models

- The models all perform equally well on the EPA data set
- The models also perform equally well when new data retains the same correlation structure
- When the correlations change, models based on inter-related variables predict less accurately than before
- The eigenfuel model is less affected by changes in the correlations



Here, we see the overall predictive ability of the 4 models summarized in terms of the median correlation coefficient between the predicted and "true" emissions response. We started with the four models performing equally well on the EPA data set. We see that:

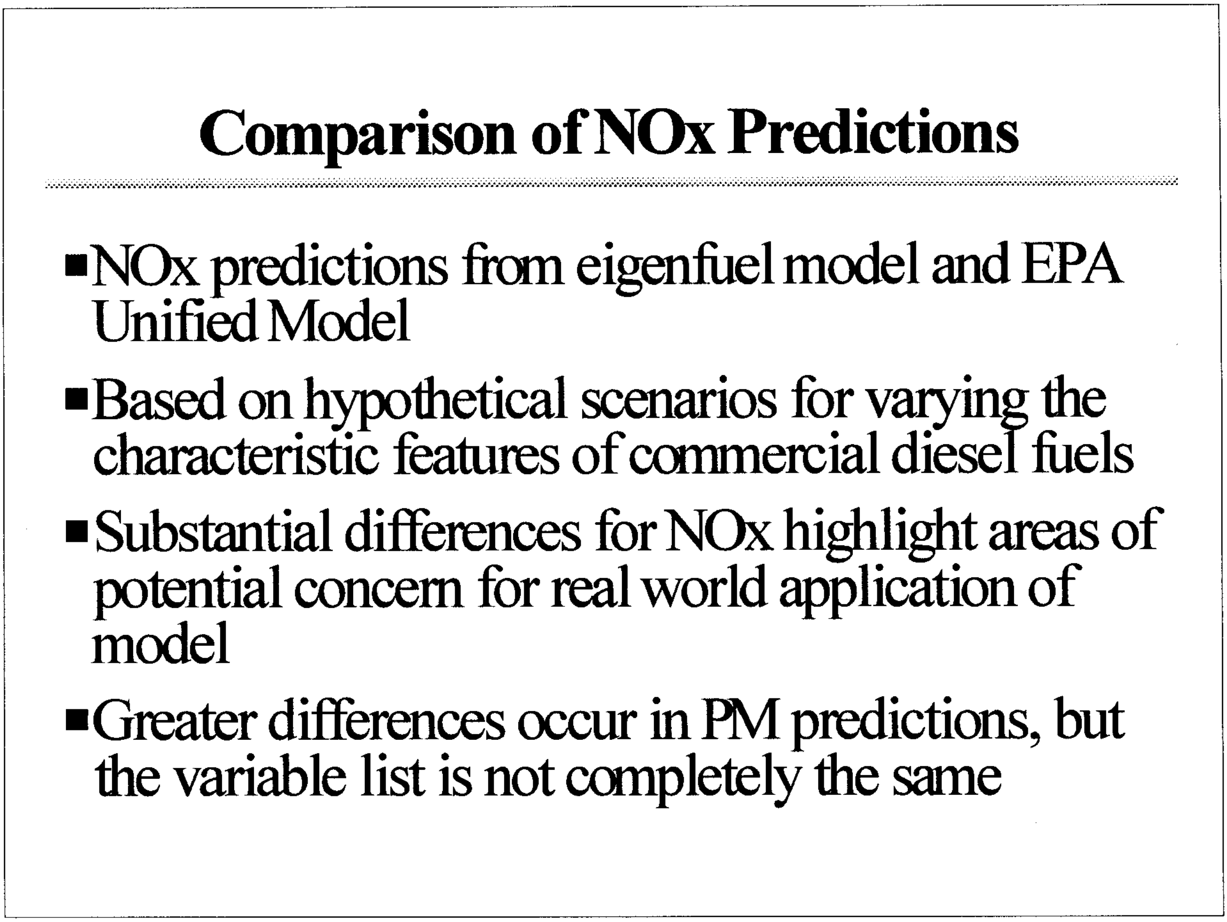
o The four models also perform equally well on new data, when that data retains the same correlation structure

o When the correlations change, the stepwise models predict less accurately than before

o The eigenfuel model is less affected by changes in the correlations and is overall a more robust basis for prediction

Basically, there is no way to avoid aliasing effects when working with correlated predictors. Any model will, to some extent at least, "tune" the coefficients to match the aliasing present in the data set. This "tuning" will falter whenever the aliasing in new data is different.

The eigenfuel model attacks this problem by eliminating the aliasing at the outset. Its coefficients are estimated in an orthogonal environment and are NOT "tuned" to any particular aliasing. It therefore predicts well in a wider range of environments.



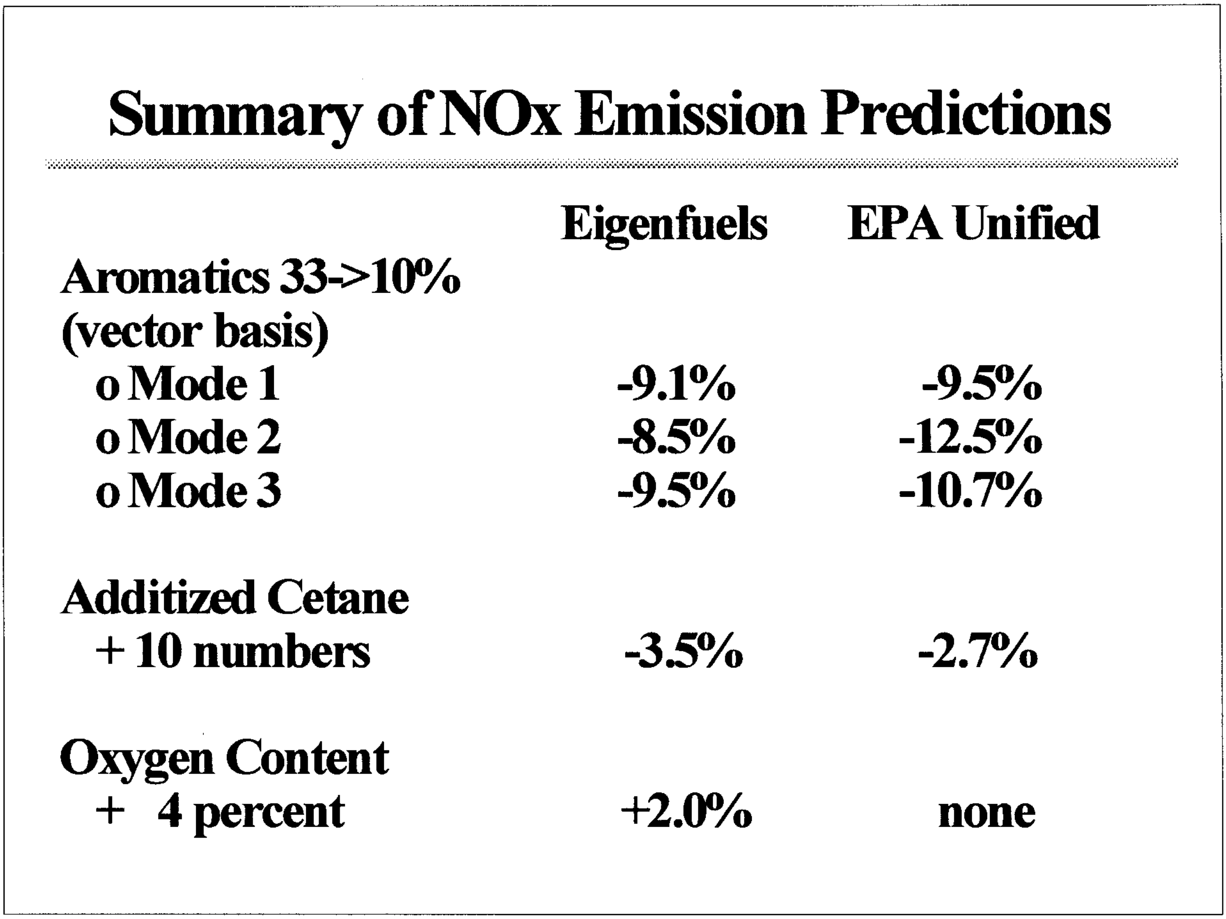
Comparison of NOx Predictions

- NOx predictions from eigenfuel model and EPA Unified Model
- Based on hypothetical scenarios for varying the characteristic features of commercial diesel fuels
- Substantial differences for NOx highlight areas of potential concern for real world application of model
- Greater differences occur in PM predictions, but the variable list is not completely the same

We have seen that the predictive ability of regressions models based on correlated predictors can deteriorate when applied to new data, while the predictive ability of the eigenfuel model is less affected.

Let us look at this issue in a different way by comparing emission predictions of the eigenfuel model and EPA's Unified Model for different hypothetical changes in diesel fuel characteristics. We will define the hypothetical fuel changes based on an analysis of the features that are characteristic of commercial diesels fuels. This is, in a sense, the acid test since what we care about is how much emission reduction we get from reformulating commercial fuels.

We cannot say that one model is right and the other is wrong, but we should be concerned if we find them to give substantially different predictions. We will base the comparison only on NOx emissions, where the two models considered the same 12 variables and the only difference is the methodology. There are actually greater differences in the emission predictions for PM, but in that case the EPA model considers an interactive term for natural cetane x cetane difference that was not available to the eigenfuel model.



Summary of NOx Emission Predictions

	Eigenfuels	EPA Unified
Aromatics 33->10% (vector basis)		
o Mode 1	-9.1%	-9.5%
o Mode 2	-8.5%	-12.5%
o Mode 3	-9.5%	-10.7%
Additized Cetane + 10 numbers	-3.5%	-2.7%
Oxygen Content + 4 percent	+2.0%	none

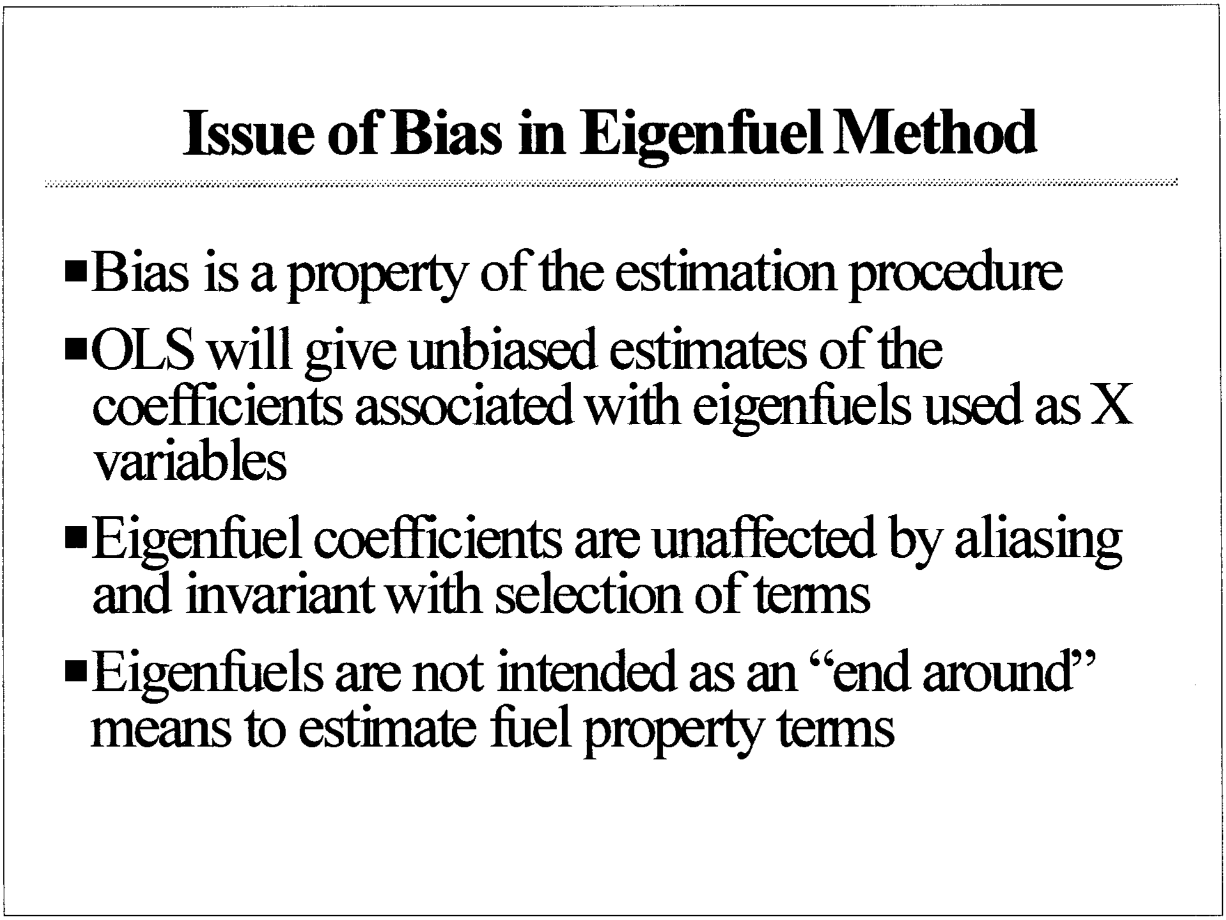
Commercial diesel fuels appear to vary in ways not fully represented in the EPA experimental data set. In particular, from an analytical perspective, there are 3 characteristically different ways (or modes) in which aromatics content can be reduced. These are defined by the differing component classes that are substituted for aromatics, each with its own effects on natural cetane, density and other properties. In comparison, there is only a single mode of aromatics variation found in the experimental fuels of the EPA database. These "modes" of aromatics reduction involve changes in a wide range of properties that are associated with aromatics, which is what we mean by "vector change".

The eigenfuel and EPA models generally agree on the NOx impact of reducing aromatics for Mode 1, but show disagreements for Modes 2 and 3. The largest disagreement is for Mode 2, where EPA's estimate is nearly half again as large. The eigenfuel model estimates about a 9 percent reduction in NOx when going from 33 to 10 percent aromatics, regardless of how this is accomplished. The EPA Unified Model gives differing values depending on mode and would appear to estimate a somewhat greater overall effect of aromatics reduction.

The models disagree on the magnitude of the additized cetane effect, with the eigenfuel model estimating a one-third greater reduction in NOx emissions than the EPA Unified Model. The models also disagree on whether oxygen content adversely impacts NOx. The eigenfuel model says that it does, while the EPA model says not, since it omits oxygen content as a predictive variable for NOx.

There are greater differences in the model predictions for PM, due both to differences in the methodology and to differences in the list of predictive variables.

Issue of Bias in Eigenfuel Models



Issue of Bias in Eigenfuel Method

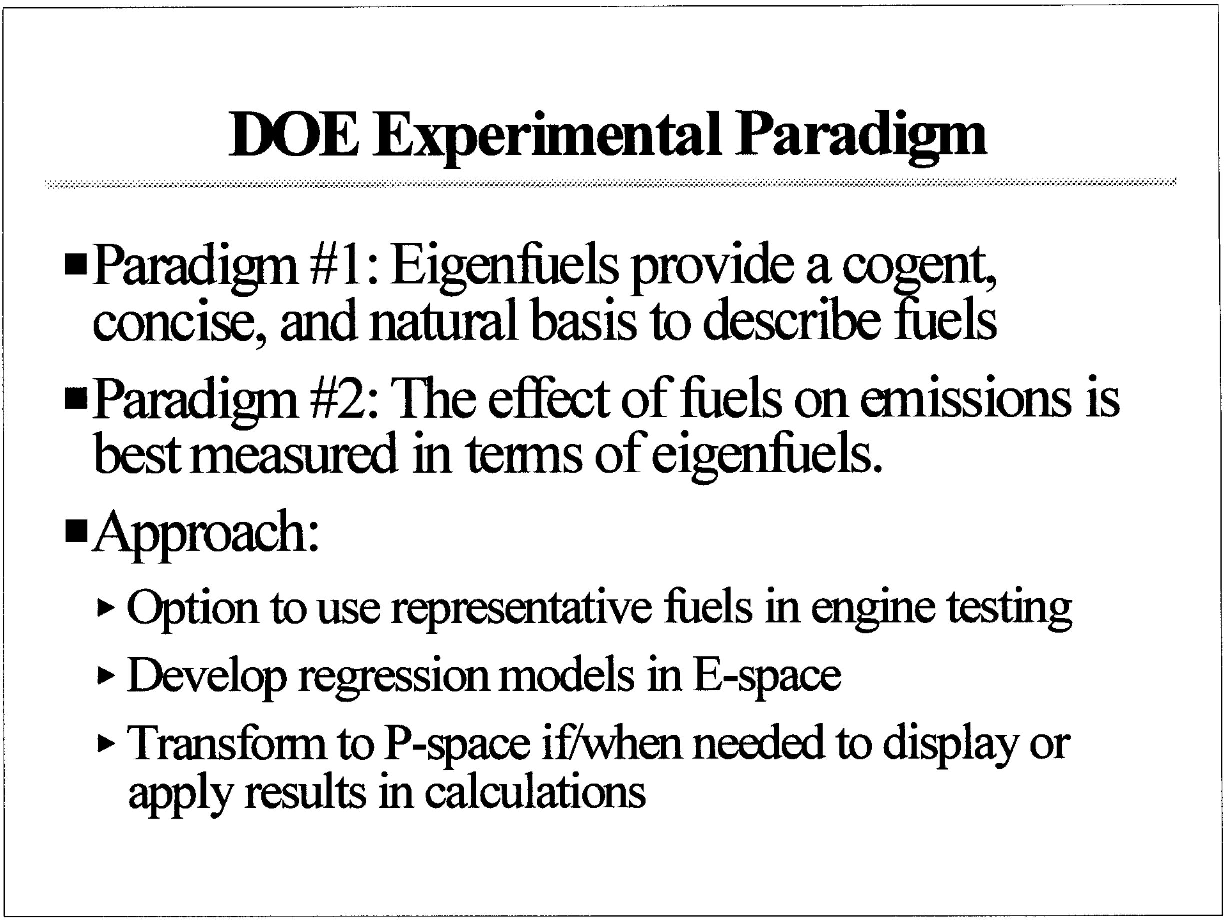
- Bias is a property of the estimation procedure
- OLS will give unbiased estimates of the coefficients associated with eigenfuels used as X variables
- Eigenfuel coefficients are unaffected by aliasing and invariant with selection of terms
- Eigenfuels are not intended as an "end around" means to estimate fuel property terms

The EPA report identifies PCR, the basis for the eigenfuel methodology, as a statistical methodology that produces biased estimates for fuel property coefficients and, therefore, does not offer a complete solution to the problems introduced by correlation.

Bias is really a property of the estimation procedure. OLS will give unbiased estimates of the coefficients associated with the X-space variables, regardless of how the X-space is defined.

When the X-space consists of the eigenfuel description of fuels, the coefficients are unbiased estimates of the response associated with the eigenvectors. In addition, the coefficients have the desirable property of being unaffected by aliasing. Therefore, the the coefficient estimates are invariant with respect to changes in the number of terms retained.

Eigenfuels are not intended as an "end round" means of estimating coefficients for the fuel property terms. Rather, we believe they are the preferrred choice of variable, because they are the only orthogonal basis that can be defined for the data set at hand. The purpose of PCR+ in our work is to estimate the response associated with the eigenvectors.



DOE Experimental Paradigm

- Paradigm #1: Eigenfuels provide a cogent, concise, and natural basis to describe fuels
- Paradigm #2: The effect of fuels on emissions is best measured in terms of eigenfuels.
- Approach:
  - Option to use representative fuels in engine testing
  - Develop regression models in E-space
  - Transform to P-space if/when needed to display or apply results in calculations

Here is the DOE paradigm for how fuels research should be conducted. We believe that eigenfuels provide a preferred basis for describing fuels and the emissions response to changing fuel characteristics.

Because of this paradigm, we advocate the following:

o The option of using commercially representative fuels in engine testing, without the artificial attempt to make fuel properties vary in unnatural ways.

o Conduct the emissions analysis in E-space, meaning using eigenvector descriptions of fuels in place of the individual properties. You should be able to use eigenfuels any where you currently use fuel properties.

o Transform your results to P-space, meaning fuel properties, as a convenience, if and when needed to display or apply the results in calculations. Caution is advised in the interpretation of the transformed equation, however.

Conclusions and Implications

Modeling with Interrelated Predictors

- Does not provide definitive guidance on fuel properties influencing emissions
- Likely to retain too few terms due to aliasing of variables and consequent inflation of standard errors
- Coefficients are "tuned" to the aliasing present in the data set
- Models that predict well for observations in the data set can not be counted on for accurate predictions elsewhere

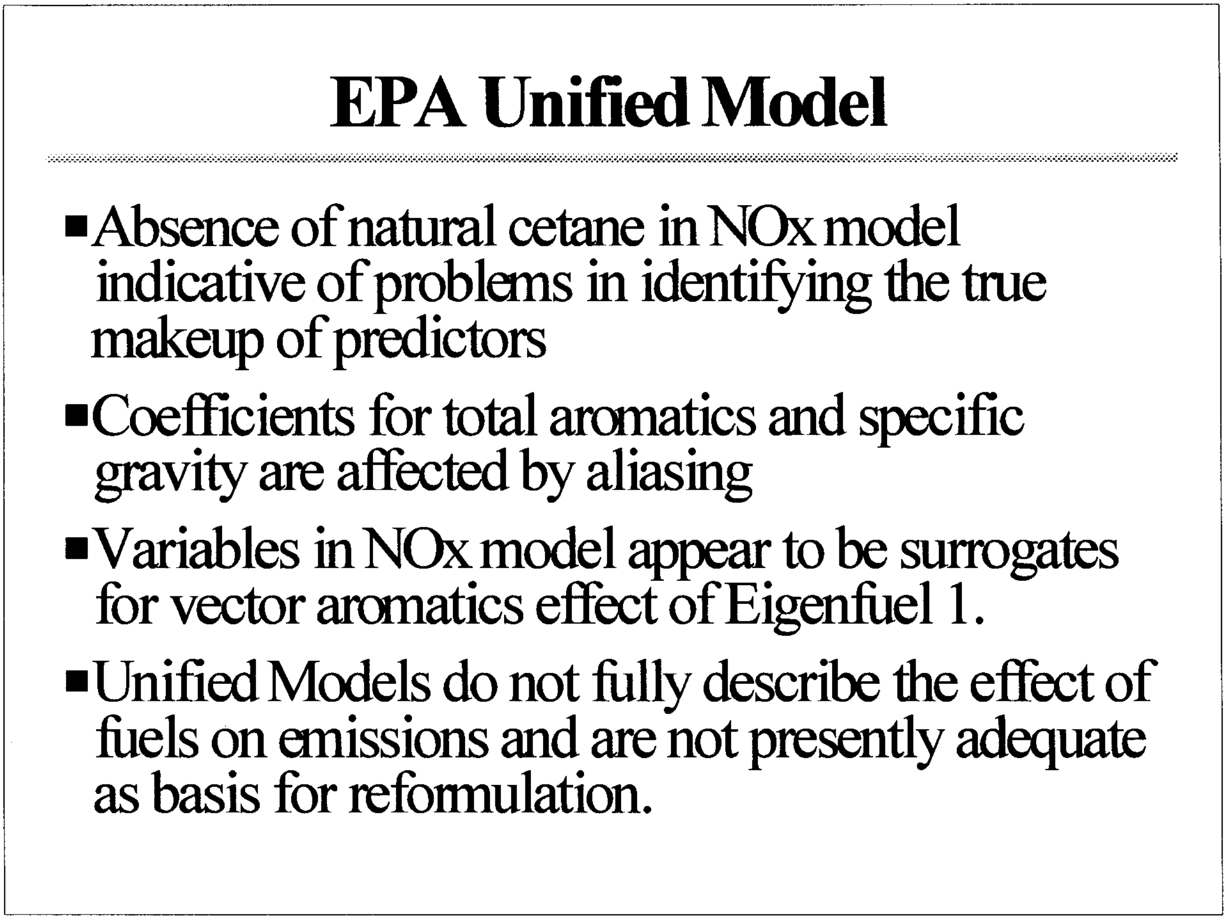
We are concerned that the current state of the art in modeling the effect of fuels on emissions falls short of the standard that needs to be met for the development of public policy in this area.

The stepwise regression methodology using inter-related variables is likely to fail the test of providing definitive guidance to refiners on the fuel properties that must be modified to reduce emissions. Too many different regression models, differing in the selection of predictor variables, are possible as outcomes of the process when the correlation among variables is appreciable.

The presence of correlations among the predictor variables also has the well-understood effect of inflating the standard errors of the estimated coefficients. When variable selection is based primarily on the t-test of statistical significance, the increased standard errors will often lead to including too few terms in the final model.

The coefficients estimated for regression models will be biased as a result of aliasing when too few predictive terms are retained. In this case, the coefficients implicitly incorporate effects associated with other variables not present in the model. The predictive ability of such models can deteriorate when applied to data having different correlation characteristics.

Modeling of fuel effects on emissions should be based on orthogonal or independent predictors to avoid the problems listed above.



EPA Unified Model

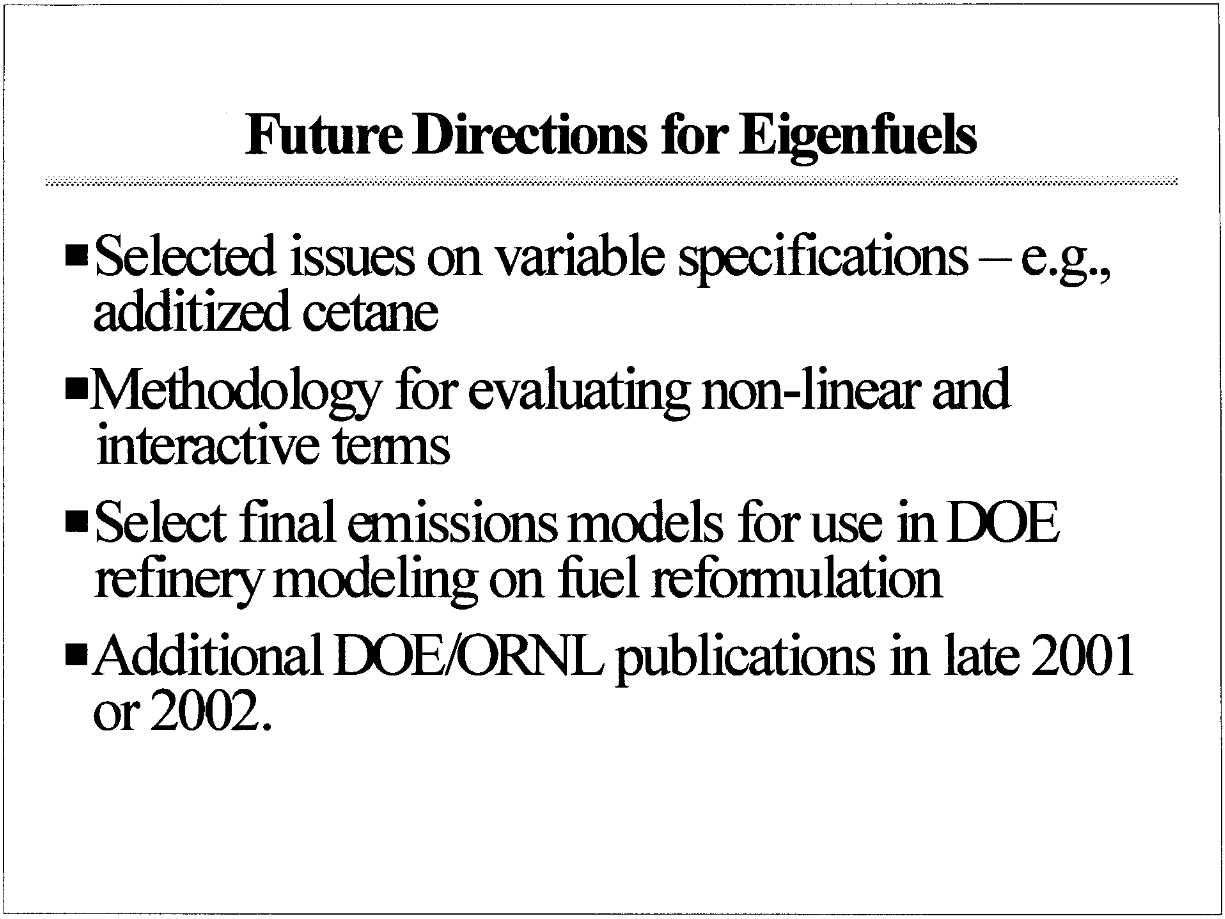
- Absence of natural cetane in NOx model indicative of problems in identifying the true makeup of predictors
- Coefficients for total aromatics and specific gravity are affected by aliasing
- Variables in NOx model appear to be surrogates for vector aromatics effect of Eigenfuel 1.
- Unified Models do not fully describe the effect of fuels on emissions and are not presently adequate as basis for reformulation.

The absence of natural cetane in EPA's unified model for NOx is, we believe, indicative of the difficulties in variable selection when operating with inter-related predictors.

We conclude that the coefficients for aromatics content and specific gravity are affected by aliasing of these variables to natural cetane and other excluded terms. These coefficients are likely to give biased estimates of the effect of varying these properties independently of other variables.

We interpret aromatics content and specific gravity as surrogate variables for the largest fuel effect identified by the eigenfuel analysis -- that is, Eigenvector 1 representing the reduction of aromatics content with its associated increase in natural cetane and decrease in specific gravity.

We expect that the stepwise estimation process has resulted in retaining too few terms in the Unified NOx and PM model to fully represent the effect of fuels. If so, the coefficient estimates are affected by aliasing and are not adequate at present to provide a sound basis for diesel fuel reformulation.



Future Directions for Eigenfuels

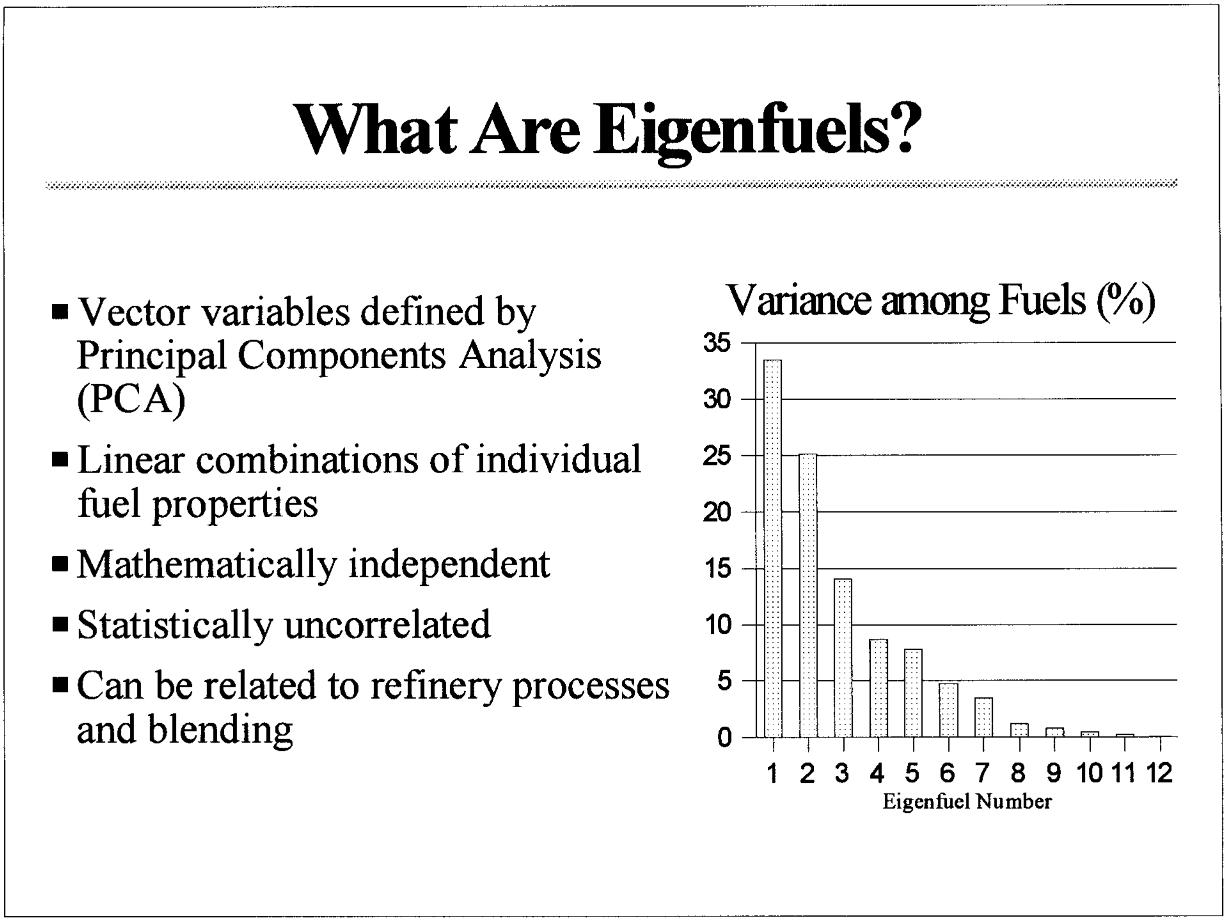
- Selected issues on variable specifications – e.g., additized cetane
- Methodology for evaluating non-linear and interactive terms
- Select final emissions models for use in DOE refinery modeling on fuel reformulation
- Additional DOE/ORNL publications in late 2001 or 2002.

This chart summarizes the directions our work on eigenfuels is likely to take. There are some methodological issues we need to address, including how best to specify some variables like additized cetane and to evaluate a full range of non-linear and interactive terms. The issue on additized cetane is whether it is more effective to include total cetane in place of cetane difference and allow the eigenvector decomposition to sort out the difference between natural and total cetane.

Having done this, our next major milestone is to select a final emissions model for use in DOE refinery analyses related to diesel fuel reformulation.

We expect to release additional publications on this work in late 2001 or early 2002.

Brief Introduction to Eigenfuel Methodology



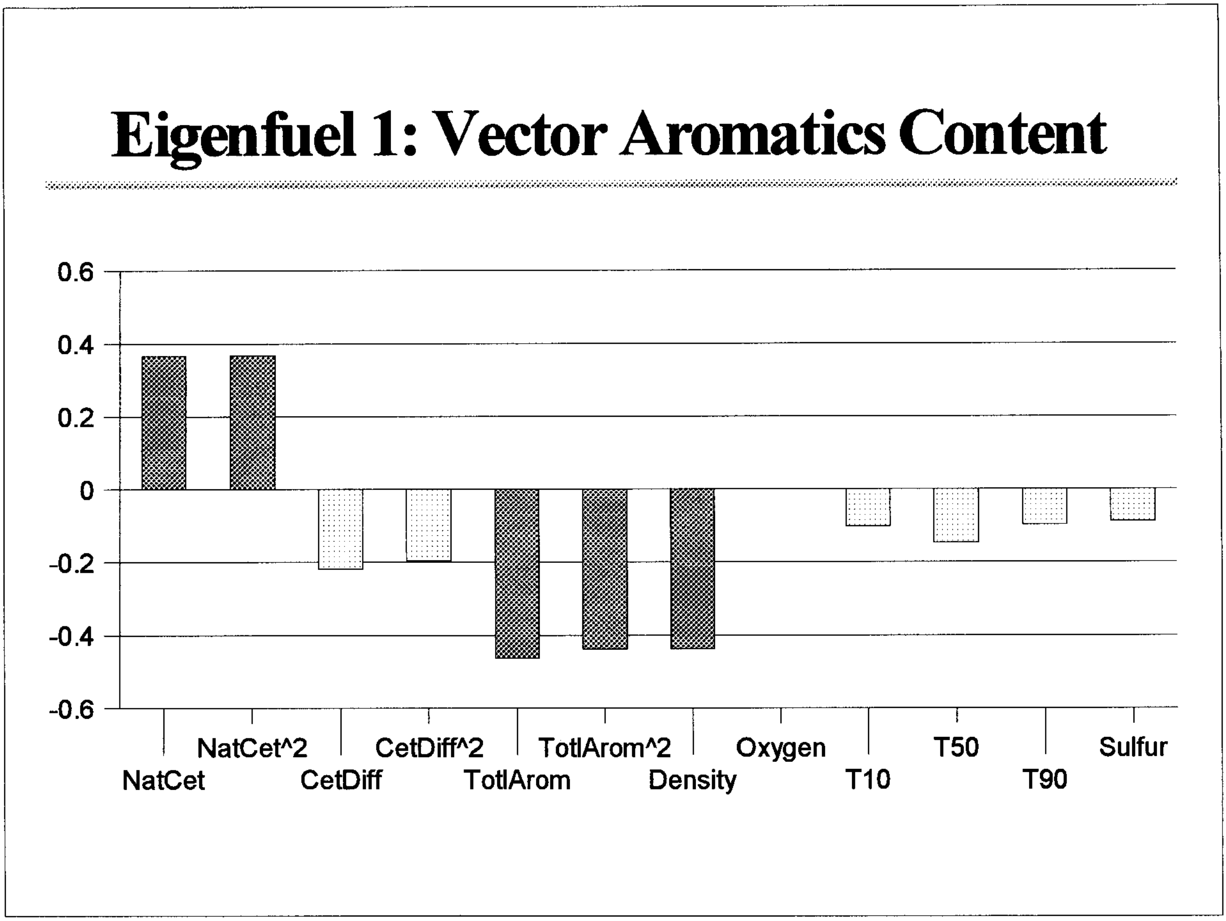
What Are Eigenfuels?

- Vector variables defined by Principal Components Analysis (PCA)
- Linear combinations of individual fuel properties
- Mathematically independent
- Statistically uncorrelated
- Can be related to refinery processes and blending

As summarized in this chart, eigenfuels are the eigenvectors defined by the Principal Components Analysis (PCA) decomposition of the correlation matrix for a data set. Each vector is expressed as a linear combination of the individual fuel properties. They are defined by the PCA procedure to be mathematically independent and statistically uncorrelated descriptors for fuels.

Eigenfuels provide an orthogonal partitioning of the variation among fuels. The bar chart shows the variance associated with each of the eigenfuels in the EPA data set. One-third of the variance is associated with the first eigenfuel feature, and one quarter with the second. Only six features are needed to explain nearly 95 percent of the differences among fuels. You can see how eigenfuels help identify the true dimensionality of the problem.

We have termed them "eigenfuels" because the vectors we have seen in both experimental and commercial fuels datasets are building blocks of fuels that have ready interpretations in terms of refining and blending processes. Allowing some time to retool one's thinking, it becomes as natural to describe fuels in terms of eigenfuels as it is to use individual fuel properties.

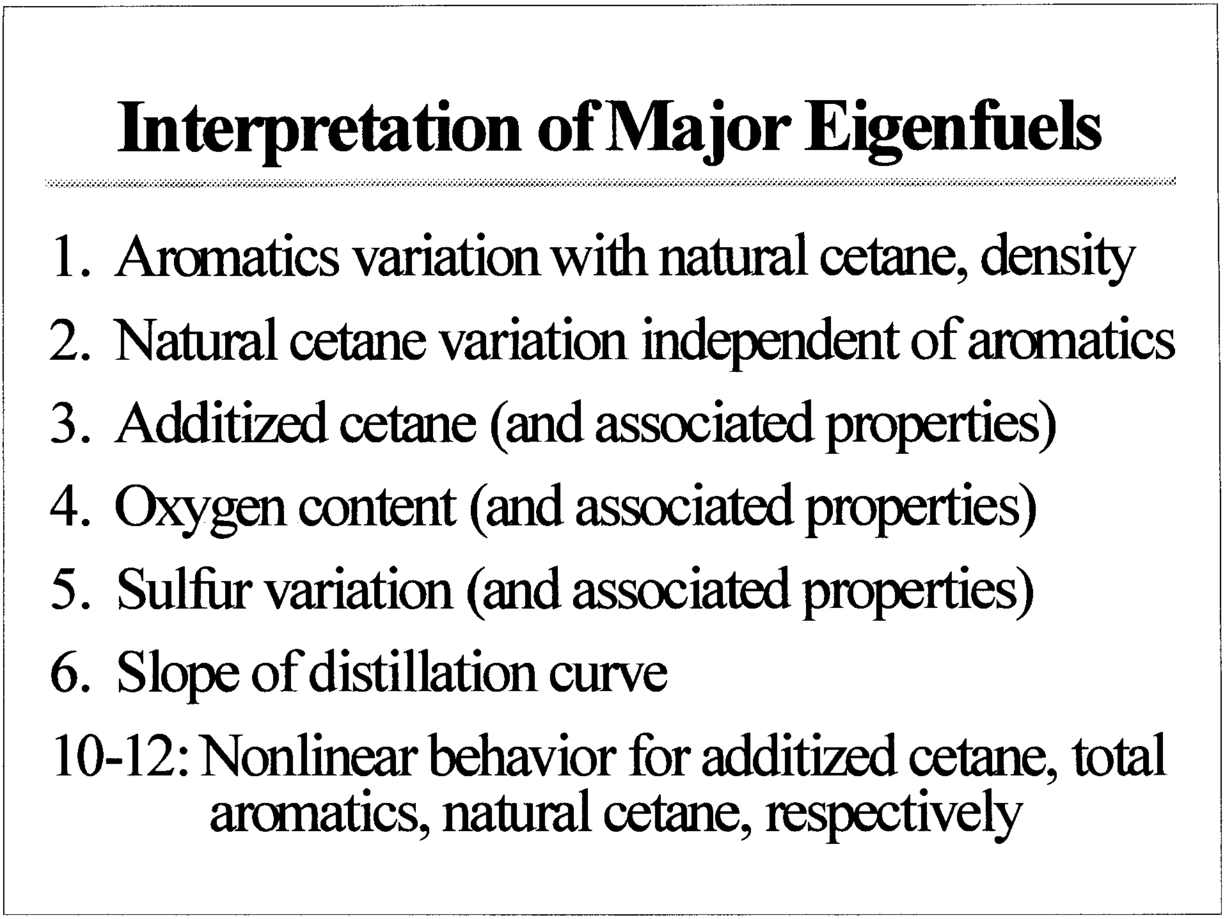


Eigenfuel 1: Vector Aromatics Content

This slide shows you graphically the composition of Eigenfuel 1. The vertical axis gives the weight for each of the 12 fuel property variables in this vector. The dark bars show you the terms that make the largest contributions to the vector.

We can "read" the vector as saying that an increase in natural cetane (X and X^2 terms equally) is associated with a decrease in total aromatics content (X and X^2 terms equally) and a decrease in density, with smaller effects on other properties. The property changes are ones that occur simultaneously whenever the amount of this eigenfuel varies.

This is the feature that varies most in the experimental fuels found in the EPA data set. A refinery-based interpretation would call this the "light cycle oil" vector.



Interpretation of Major Eigenfuels

1. Aromatics variation with natural cetane, density
2. Natural cetane variation independent of aromatics
3. Additized cetane (and associated properties)
4. Oxygen content (and associated properties)
5. Sulfur variation (and associated properties)
6. Slope of distillation curve

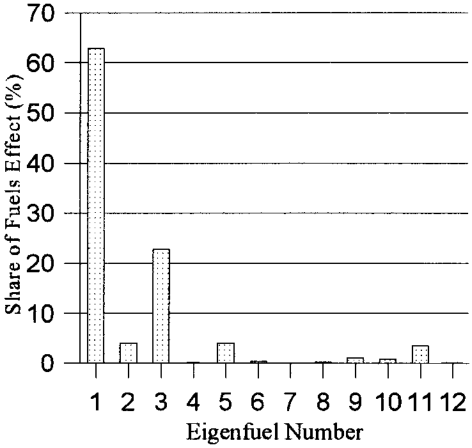
10-12: Nonlinear behavior for additized cetane, total aromatics, natural cetane, respectively

Here are interpretations for the major eigenfuels. The first five vectors are ones you might logically expect to find in a data set developed to test for effects on emissions. The sixth feature -- the slope of the distillation curve -- seems related to controlling flash and pour points to commercial specifications. Nonlinear behavior for additized cetane, total aromatics, and natural cetane are represented in eigenfuels 10, 11, and 12.

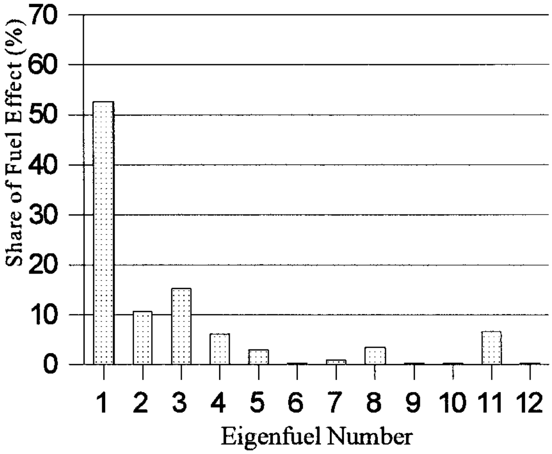
The "bottom line" to this slide is that eigenfuels offer a cogent and concise method for describing fuels that reveals how fuels have been formed and can be a much more natural basis than the individual properties.

Eigenfuel Impact on Emissions

NOx Emissions



PM Emissions



Real fuels can easily be expressed in terms of eigenfuels, and the coefficients of that expression can then used predictors in regression analysis for NOx and PM emissions. In this example, effects related to individual engines have first been removed from the data, leaving us with the effects of fuels and unexplained variation. The regression is of the form log(emissions) = f(eigenfuels).

The vertical axis is the percent contribution of each eigenfuel to the fuels-related model SS. This contribution depends on both strength (magnitude) of the eigenfuel effect and how much each eigenfuel was varied in the test set. A small share here does not necessarily mean that an eigenfuel is unimportant, since it might vary much more in another dataset.

We see that eigenfuel number 1 -- the aromatics, natural cetane, and density vector -- has the single largest effect on both log(NOx) and log(PM). From the perspective of eigenfuels, the effects are related to the joint variation in aromatics, natural cetane, and density. They cannot be ascribed to any one fuel property in isolation from the others.

For NOx, we find 8 terms to be statistically significant: 1, 2, 3, 5, 6, 9, 10, 11. After the aromatics vector, additized cetane (3) is next most important, with natural cetane independent of aromatics (2), sulfur content (5), and non-linear aromatics (11) making smaller and nearly equal contributions.

For PM, we find 7 terms to be statistically significant: 1, 2, 3, 4, 5, 8, 11. After the aromatics vector, additized cetane (3) is next most important, followed by natural cetane independent of aromatics (2), oxygen content (4) and non-linear aromatics (11).

Note also that the eigenfuel terms that are significant in explaining NOx or PM emissions are not the same as the ones most important for describing the variation among fuels. This is an example of why one cannot prune the eigenvector terms before the fact based only on what is important in describing fuels.